Consistent Estimation, Variable Selection, and Forecasting in Factor-Augmented VAR Models*

John C. Chao[1], Yang Liu[2], Kaiwen Qiu[3], and Norman R. Swanson[3]

[1]University of Maryland, [2]Towson University, and [3]Rutgers University

September 8, 2025

Abstract

We introduce a completely consistent method for variable selection with high dimensional datasets. The method is presented in a framework where latent factors are estimated for the purpose of dimension reduction, and is meant to serve as a complement to extant methods. We argue that the method is of particular interest in empirical settings where there may be many irrelevant predictor variables. The reason for this is that situations where there are "too many" irrelevant variables can lead to inconsistent factor estimates. Interestingly, our method yields a consistent estimate of the number of such irrelevant variables, which can aid the applied practitioner in assessing the strength of the underlying factor structure for a particular application. We also show that when factors constructed using our variable selection method are inputted into the forecast equations implied by a factor augmented vector autoregressive (FAVAR) model for the purpose of forecasting, the conditional mean forecast equations can be consistently estimated. Monte Carlo results are presented indicating that the variable selection method performs well in finite samples. The paper also contains two empirical illustrations, where we compare forecasts constructed from factor-augmented forecast equations using our variable selection procedure with two alternative methods for factor construction - the conventional PCA procedure, which does not pre-screen the variables, and a hard thresholding method that is widely used in the empirical literature. Overall, we find that our method to outperform both of these alternative procedures in a majority of the cases that we study across different target variables, forecast horizons, and data window specifications (i.e., recursive or rolling).

Keywords: Factor analysis, factor augmented vector autoregression, forecasting, moderate deviation, principal components, self-normalization, variable selection.

JEL Classification: C32, C33, C38, C52, C53, C55.

*John C. Chao, Department of Economics, 7343 Preinkert Drive, University of Maryland, chao@econ.umd.edu.
Kaiwen Qiu, Department of Economics, 9500 Hamilton Street, Rutgers University, kq60@economics.rutgers.edu.
Norman R. Swanson, Department of Economics, 9500 Hamilton Street, Rutgers University, nswanson@econ.rutgers.edu.
Yang Liu, Department of Economics, 8000 York Rd, Towson, MD 21252, Towson University, yliu@towson.edu.
The authors are grateful to Matteo Barigozzi, Rong Chen, Harold Chiang, Domenico Giannone, Bruce Hansen, Cheng Hsiao, Yuan Liao, Jack Porter, Xiaoxia Shi, Minchul Shin, Xiye Yang, Peter Zadrosny and participants at the 2022 International Association of Applied Econometrics conference, the 2023 AiE Conference in Honor of Joon Park, the 2024 Kansas Econometrics Workshop, the 2023 Rochester Conference in Econometrics, the 2024 Annual Conference of the Society for Economic Measurement, the Federal Reserve Bank of Philadelphia, and the univeristies of Glasgow, Riverside, and Wisconsin-Madison for useful comments received on earlier versions of this paper. Chao thanks the University of Maryland for research support.

1 Introduction

In this paper, we propose a simple to implement and completely consistent method for variable selection when estimating factors for use in dimension reduction and factor augmented vector autoregression (FAVAR) forecasting with high dimensional datasets. Our method is meant to add to variable selection methods available to empirical practitioners that have been introduced in recent papers, including those due to Bair, Hastie, Paul, and Tibshirani (2006), Bai and Ng (2008), Giglio, Xiu, and Zhang (2023a,b), among others. In addition, our method builds on previous supervised machine learning approaches that are now widely used in the literature, including, and not limited to principal components analysis (PCA), sparse, and supervised PCA (see, e.g. Zou, Hastie, and Tibshirani (2006), Barshan, Ghodsi, Azimifar, and Jahromi (2011), Carrasco and Rossi (2016), and Fan, Ke and Liao (2021)); bagging, boosting, and random forest (see, e.g. Breiman (1996), Freund and Schapire (1997), Breiman (2001), Lee and Yang (2006), Lee, Ullah, and Wang (2020), and the references cited therein); and regression methods such as the elastic net, garrote, and lasso (see e.g. Breiman (1995), Tibshirani (1996), Zou and Hastie (2005), Kim and Swanson (2014), Belloni, Chernozhukov, and Wang (2014), and the references cited therein).

The variable selection procedure introduced here seeks to identify and eliminate those irrelevant variables which do not load on any of the underlying factors so that only relevant variables which contain information about at least one of the latent factors are used in factor estimation.[1] This is of importance because the use of irrelevant variables in extracting the latent factors could lead to less accuracy since these variables contribute only noise but not signal to the factor estimation process. Although variable selection procedures for factor estimation have also been studied in the well-known paper by Bair, Hastie, Paul, and Tibshirani (2006) on supervised principal component methods and in some interesting recent papers by Giglio, Xiu, and Zhang (2023a,b), a notable difference between our selection method and those proposed in these other papers is our use of a self-normalized statistic. A key attribute of self-normalized statistics is that their tail behavior can be better approximated over a wider range, using moderate deviation results, than statistics which are not self-normalized, as we will explain in greater details in Section 2 of the paper. This, in turn, allows us to specify our decision rule in such a way so that our procedure will be completely

[1] Although we interpret our variable selection procedure primarily as a procedure which selects variables on the basis of relevance, it should be noted that in a forecasting context the procedure considered here can also be useful for assessing whether a particular variable has predictive content for the target variable of interest. Please see Remark 2.2 for a detailed discussion about the close relationship between the relevance and the predictive content of a variable in the context of a FAVAR model and how a score statistic can be useful in assessing both.

consistent in the sense that the probability of Type I and Type II errors will both go to zero as sample sizes approach infinity[2].

An important added value of our completely consistent variable selection procedure is that it enables us to construct a consistent estimator $\widehat{N}_1$ of N_1 (the number of relevant variables) in the sense that $\widehat{N}_1/N_1 \xrightarrow{p} 1$. As explained in Section 2 of this paper, since N_1 is itself not directly observable, having a consistent estimator $\widehat{N}_1$ provides empirical researchers with a useful diagnostic statistic which can help them assess the overall pervasiveness of the factors in empirical applications. As discussed in Section 2, consistent estimation of N_1 will not be possible if one does not have a selection method where the probability of a Type II error approaches zero asymptotically. In addition, we will also not be able to consistently estimate N_1 if the probability of a Type I error is not controlled to vanish asymptotically, except in the special case where N_2, the number of irrelevant variables, is negligible relative to N_1 (i.e., the case where $N_2/N_1 \to 0$). However, if $N_2/N_1 \to 0$ then forecast results based on use of our procedure should not be much different from forecast results obtained from the use of conventional PCA (where no variable pre-screening is conducted). This does not seem to be the case, in particular, for the FRED-MD dataset that we examine in an empirical illustration (see Section 5 for further details).[3]

To properly control the probability of a Type I error in our setup, we leverage on some important advances in moderate deviation results for weakly dependent processes obtained recently by Chen, Shao, Wu, and Xu (2016). In the context of the moderate deviation theory used here, a further advantage of self-normalized statistics is that relative to their non-self-normalized counterparts, statistics which are self-normalized are more able to accommodate situations where the underlying distribution of the data may have thicker tails. Hence, moderate deviation results for self-normalized statistics require weaker moment conditions than statistics which are not self-normalized.

In addition to proposing a new variable selection method and showing its complete consistency, we also make a number of contributions to the methodology of carrying out forecasting in a dynamic factor-augmented modeling framework. More specifically, within a general FAVAR setup, which allows time series forecasts to be made using information sets much richer than that used in traditional VAR models, we provide an easy-to-implement formula for the post-variable-selection principal component estimator of the vector of factors. We then show that this post-variable-selection factor estimator can consistently estimate the true factors up to an invertible matrix

[2]Here, we take Type I error to be the error that an irrelevant variable is falsely selected as a relevant variable, whereas Type II error is the error of misclassifying a relevant variable as being irrelevant.

[3]A formal proof of the consistency of $\widehat{N}_1$ is given in part (a) of Lemma C-15 of the Technical Appendix. In addition, we also provide in Section 2 of the paper some intuitive discussion about why having both the probability of a Type I error and that of a Type II error vanish asymptotically is important for the consistency of $\widehat{N}_1$. Please see, in particular, Example 2 given in Remark 2.3.

transformation even if we do not impose the kind of normalization conditions on the factors and the factor loadings that are typically made in the literature; see, for example, Stock and Watson (2002a). Moreover, we explicitly derive a closed-form representation for the system of h-step ahead forecasting equations implied by a FAVAR model and show that, by inserting our post-variable-selection factor estimates into the h-step ahead forecasting equations, we can consistently estimate the conditional mean function of the said equations; and this is true even if we do not make strong enough identifying assumptions so that the factors can only be consistently estimated up to an invertible matrix transformation.

Besides our theoretical results, we also present Monte Carlo results which indicate that the finite sample performance of our variable selection procedure is in accord with the results of our asymptotic analysis. In particular, when the sample sizes are large, such as the case where $T = 600$ and $N = 1000$, then results of our simulation study show that both Type I and Type II error rates are very close to zero. Moreover, even in the smaller sample case where $T = 100$ and $N = 100$, the Type I and Type II error rates are usually less than 0.05, and are often much smaller than that. In addition we carry out two small forecasting exercises to illustrate the empirical relevance of our procedure. The first uses a variety of macroeconomic variables from the well-known FRED-MD database and the second employs data from an updated version of the GVAR dataset previously studied in Dees, di Mauro, Pesaran, and Smith (2007) and Pesaran, Schuermann, and Smith (2009). In both empirical illustrations, we compare forecasts constructed from factor-augmented forecast equations using our variable selection procedure with two alternative methods for factor construction (the conventional PCA procedure, which does not pre-screen the variables, and a hard thresholding method that is widely used in the empirical literature). Overall, we find our method to outperform both of these alternative procedures in a majority of the cases that we study across different target variables, forecast horizons, and data window specifications (i.e., recursive or rolling). We believe that our results indicate the potential usefulness of our method in empirical applications.

The rest of the paper is organized as follows. In Section 2, we discuss the FAVAR model and the assumptions that we impose on this model. We also describe our variable selection procedure and provide theoretical results establishing the complete consistency of the procedure. Section 3 provides theoretical results on the consistent estimation of latent factors, up to an invertible matrix transformation, as well as results on the consistent estimation of the h-step ahead predictor, based on the FAVAR model. Section 4 presents the results of a promising Monte Carlo study on the finite sample performance of our variable selection method, and makes recommendations regarding the calibration of the tuning parameter used in said method. Section 5 presents the results of two empirical applications comparing forecast results based on our method with those obtained from

PCA and the hard thresholding method of Bai and Ng (2008). Finally, Section 6 offers concluding remarks. Proofs of the main theorems as well as additional supporting lemmas are given in a separate Technical Appendix.[4] The Technical Appendix is organized into three subappendices. Appendix A provides proofs of the main theorems. Appendix B contains proofs of supporting lemmas, used primarily in the proofs of Theorems 1 and 2, and Appendix C contains the proofs of supporting lemmas used primarily in the proofs of Theorems 3 and 4.

Before proceeding, we first say a few words about some of the notation used in this paper. Throughout, let $\lambda_{(j)}(A)$, $\lambda_{\max}(A)$, $\lambda_{\min}(A)$, and $tr(A)$ denote, respectively, the j^{th} largest eigenvalue, the maximal eigenvalue, the minimal eigenvalue, and the trace of a square matrix A. Similarly, let $\sigma_{(j)}(B)$, $\sigma_{\max}(B)$, and $\sigma_{\min}(B)$ denote, respectively, the j^{th} largest singular value, the maximal singular value, and the minimal singular value of a matrix B, which is not restricted to be a square matrix. In addition, let $\|a\|_2$ denote the usual Euclidean norm when applied to a (finite-dimensional) vector a. Also, for a matrix A, $\|A\|_2 \equiv \max\left\{\sqrt{\lambda(A'A)} : \lambda(A'A) \text{ is an eigenvalue of } A'A\right\}$ denotes the matrix spectral norm, and $\|A\|_F \equiv \sqrt{tr\{A'A\}}$ denotes the Frobenius norm. For two random variables X and Y, write $X \sim Y$, if $X/Y = O_p(1)$ and $Y/X = O_p(1)$. Furthermore, let $|z|$ denote the absolute value or the modulus of the number z; let $\lfloor \cdot \rfloor$ denote the floor function, so that, for a real number x, $\lfloor x \rfloor$ gives the largest integer that is less than or equal to x; let $\lceil \cdot \rceil$ denote the ceiling function, so that, for a real number x, $\lceil x \rceil$ gives the smallest integer that is greater than or equal to x; and let $\iota_p = (1, 1, ..., 1)'$ denote a $p \times 1$ vector of ones. Finally, the abbreviation w.p.a.1 stands for "with probability approaching one".

2 Model, Assumptions, and Variable Selection in High Dimensions

Consider the following p^{th}-order factor-augmented vector autoregression (FAVAR):

$$W_{t+1} = \mu + A_1 W_t + \cdots + A_p W_{t-p+1} + \varepsilon_{t+1}, \tag{1}$$

[4] The technical appendix is posted online at: http://econweb.rutgers.edu/nswanson/papers/ and also at http://econweb.umd.edu/~chao/Research/research.html

where

$$\underset{(d+K)\times 1}{W_{t+1}} = \begin{pmatrix} \underset{d\times 1}{Y_{t+1}} \\ \underset{K\times 1}{F_{t+1}} \end{pmatrix}, \underset{(d+K)\times 1}{\varepsilon_{t+1}} = \begin{pmatrix} \underset{d\times 1}{\varepsilon^{Y}_{t+1}} \\ \underset{K\times 1}{\varepsilon^{F}_{t+1}} \end{pmatrix}, \underset{(d+K)\times 1}{\mu} = \begin{pmatrix} \underset{d\times 1}{\mu_{Y}} \\ \underset{K\times 1}{\mu_{F}} \end{pmatrix}, \text{ and}$$

$$\underset{(d+K)\times(d+K)}{A_g} = \begin{pmatrix} \underset{d\times d}{A_{YY,g}} & \underset{d\times K}{A_{YF,g}} \\ \underset{K\times d}{A_{FY,g}} & \underset{K\times K}{A_{FF,g}} \end{pmatrix}, \text{ for } g = 1, ..., p.$$

Here, Y_t denotes the vector of observable economic variables, and F_t is a vector of unobserved (latent) factors. In our analysis of this model, it will often be convenient to rewrite the FAVAR in several alternative forms, such as when making assumptions used in the sequel. We thus briefly outline two alternative representations of the above model. First, it is easy to see that the system of equations given in (1) can be written in the form:

$$Y_{t+1} = \mu_Y + A_{YY}\underline{Y}_t + A_{YF}\underline{F}_t + \varepsilon^{Y}_{t+1}, \tag{2}$$

$$F_{t+1} = \mu_F + A_{FY}\underline{Y}_t + A_{FF}\underline{F}_t + \varepsilon^{F}_{t+1}, \tag{3}$$

where

$$\underset{d\times dp}{A_{YY}} = \begin{pmatrix} A_{YY,1} & A_{YY,2} & \cdots & A_{YY,p} \end{pmatrix}, \underset{d\times Kp}{A_{YF}} = \begin{pmatrix} A_{YF,1} & A_{YF,2} & \cdots & A_{YF,p} \end{pmatrix},$$

$$\underset{K\times dp}{A_{FY}} = \begin{pmatrix} A_{FY,1} & A_{FY,2} & \cdots & A_{FY,p} \end{pmatrix}, \underset{K\times Kp}{A_{FF}} = \begin{pmatrix} A_{FF,1} & A_{FF,2} & \cdots & A_{FF,p} \end{pmatrix},$$

and where

$$\underset{dp\times 1}{\underline{Y}_t} = \begin{pmatrix} Y_t \\ Y_{t-1} \\ \vdots \\ Y_{t-p+1} \end{pmatrix}, \text{ and } \underset{Kp\times 1}{\underline{F}_t} = \begin{pmatrix} F_t \\ F_{t-1} \\ \vdots \\ F_{t-p+1} \end{pmatrix}. \tag{4}$$

Another useful representation of the FAVAR model is the so-called companion form, wherein the p^{th}-order model given in expression (1) is written in terms of a first-order model:

$$\underset{(d+K)p\times 1}{\underline{W}_t} = \alpha + A\underline{W}_{t-1} + E_t,$$

where $\underline{W}_t = \begin{pmatrix} W_t' & W_{t-1}' & \cdots & W_{t-p+2}' & W_{t-p+1}' \end{pmatrix}'$ and where

$$\alpha = \begin{pmatrix} \mu \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}, A = \begin{pmatrix} A_1 & A_2 & \cdots & A_{p-1} & A_p \\ I_{d+K} & 0 & \cdots & 0 & 0 \\ 0 & I_{d+K} & \ddots & \vdots & 0 \\ \vdots & \ddots & \ddots & 0 & \vdots \\ 0 & \cdots & 0 & I_{d+K} & 0 \end{pmatrix}, \text{ and } E_t = \begin{pmatrix} \varepsilon_t \\ 0 \\ \vdots \\ 0 \\ 0 \end{pmatrix}. \tag{5}$$

This companion form is convenient for establishing certain moment conditions on $\underline{Y}_t$ and $\underline{F}_t$, given a moment condition on ε_t, and for establishing certain mixing properties of the FAVAR model, as shown in the proofs of Lemmas B-5 and Lemma B-11 given in Appendix B.

In addition to observations on Y_t, suppose that the data set available to researchers includes a vector of time series variables which are related to the unobserved factors in the following manner:

$$\underset{N\times 1}{Z_t} = \Gamma \underline{F}_t + u_t, \tag{6}$$

where the properties of u_t are given in Assumptions 2-3 and 2-4, below. Now, assume that not all components of Z_t provide useful information for estimating the unobserved vector, $\underline{F}_t$, so that the $N \times Kp$ parameter matrix Γ may have some rows whose elements are all zero. More precisely, let the $1 \times Kp$ vector, γ_i', denote the i^{th} row of Γ, and assume that the rows of the matrix Γ can be divided into two classes:

$$H = \{k \in \{1, ..., N\} : \gamma_k = 0\} \text{ and} \tag{7}$$

$$H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}. \tag{8}$$

In this case, there exists a permutation matrix $\mathcal{P}$ such that $\mathcal{P}Z_t = \begin{pmatrix} Z_t^{(1)\prime} & Z_t^{(2)\prime} \end{pmatrix}'$, where

$$\underset{N_1\times 1}{Z_t^{(1)}} = \Gamma_1 \underline{F}_t + u_t^{(1)} \tag{9}$$

$$\underset{N_2\times 1}{Z_t^{(2)}} = u_t^{(2)}. \tag{10}$$

The above representation suggests that the components of $Z_t^{(1)}$ can be interpreted as some sort of "information" variables, as the information that they supply will be helpful in estimating $\underline{F}_t$. On the other hand, for the purpose of factor estimation, the components of the subvector $Z_t^{(2)}$ are pure "noise" variables, as they do not load on the underlying factors and only add noise if

they are included in the factor estimation process. An empirical researcher will often not have prior knowledge as to which variables are elements of $Z_t^{(1)}$ and which are elements of $Z_t^{(2)}$. This underscores the potential usefulness for a variable selection procedure which will allow us to properly identify the components of of $Z_t^{(1)}$ and to use only these variables when we try to estimate $\underline{F}_t$. If we unknowingly include too many components of $Z_t^{(2)}$ in the estimation process, then inconsistency in factor estimation can result, as shown in Theorem 2.1 of Chao, Qiu, and Swanson (2023).[5]

To provide a variable selection procedure with provable guarantees, we must first specify a number of conditions on the FAVAR model defined above.

Assumption 2-1: Suppose that:

$$\det\left\{I_{(d+K)} - A_1 z - \cdots - A_p z^p\right\} = 0, \text{ implies that } |z| > 1. \tag{11}$$

Assumption 2-2: Let ε_t satisfy the following set of conditions: (a) $\{\varepsilon_t\}$ is an independent sequence of random vectors with $E[\varepsilon_t] = 0\ \forall t$; (b) there exists a positive constant C such that $\sup_t E\|\varepsilon_t\|_2^6 \leq C < \infty$; (c) ε_t admits a density g_{ε_t} such that, for some positive constant $M < \infty, \sup_t \int |g_{\varepsilon_t}(\upsilon - u) - g_{\varepsilon_t}(\upsilon)|\, d\varepsilon \leq M|u|$, whenever $|u| \leq \overline{\kappa}$ for some constant $\overline{\kappa} > 0$; and (d) there exists a constant $\underline{C} > 0$ such that $\inf_t \lambda_{\min}\{E[\varepsilon_t \varepsilon_t']\} \geq \underline{C} > 0$.

Assumption 2-3: Let $u_{i,t}$ be the i^{th} element of the error vector u_t in expression (6), and we assume that it satisfies the following conditions: (a) $E[u_{i,t}] = 0$ for all i and t; (b) there exists a positive constant $\overline{C}$ such that $\sup_{i,t} E|u_{i,t}|^7 \leq \overline{C} < \infty$, and there exists a constant $\underline{C} > 0$ such that $\inf_{i,t} E\left[u_{i,t}^2\right] \geq \underline{C}$; (c) define $\mathcal{F}_{i,-\infty}^t = \sigma(...., u_{i,t-2}, u_{i,t-1}, u_t)$, $\mathcal{F}_{i,t+m}^{\infty} = \sigma(u_{i,t+m}, u_{i,t+m+1}, u_{i,t+m+2}, ...)$, and

$$\beta_i(m) = \sup_t E\left[\sup\left\{\left|P\left(B|\mathcal{F}_{i,-\infty}^t\right) - P(B)\right| : B \in \mathcal{F}_{i,t+m}^{\infty}\right\}\right].$$

Assume that there exist constants $a_1 > 0$ and $a_2 > 0$ such that

$$\beta_i(m) \leq a_1 \exp\{-a_2 m\}, \text{ for all } i;$$

and (d) there exists a positive constant C such that $\sup_t\left(\frac{1}{N_1}\sum_{i\in H^c}\sum_{k\in H^c}|E[u_{i,t}u_{k,t}]|\right) \leq C < \infty$ for every positive integer N_1, where H^c is defined in expression (8) above.

Assumption 2-4: ε_t and $u_{i,s}$ are independent, for all i, t, and s.

Assumption 2-5: There exists a positive constant $\overline{C}$, such that $\sup_{i\in H^c}\|\gamma_i\|_2 \leq \overline{C} < \infty$ and $\|\mu\|_2 \leq \overline{C} < \infty$, where $\mu = (\mu_Y', \mu_F')'$.

[5] Chao, Qiu, and Swanson (2023) is a not-for-publication working paper and can be found at http://econweb.umd.edu/~chao/Research/research.html Note that some of the results in the current paper draw on results contained in Chao, Qiu, and Swanson (2023).

Assumption 2-6: There exists a positive constant $\overline{C}$, such that:

$$0 < \frac{1}{\overline{C}} \leq \lambda_{\min}\left(\frac{\Gamma'\Gamma}{N_1}\right) \leq \lambda_{\max}\left(\frac{\Gamma'\Gamma}{N_1}\right) \leq \overline{C} < \infty \text{ for all } N_1,\, N_2 \text{ sufficiently large,}$$

where N_1 is the number of components of the subvector $Z_t^{(1)}$ (i.e., the number of relevant variables) and N_2 is the number of components of the subvector $Z_t^{(2)}$ (i.e., the number of irrelevant variables), as previously defined in expressions (9) and (10).

Assumption 2-7: Let A be as defined in expression (5) above, and let the eigenvalues of the matrix $I_{(d+K)p} - A$ be sorted so that:

$$\left|\lambda_{(1)}\left(I_{(d+K)p} - A\right)\right| \geq \left|\lambda_{(2)}\left(I_{(d+K)p} - A\right)\right| \geq \cdots \geq \left|\lambda_{((d+K)p)}\left(I_{(d+K)p} - A\right)\right| = \overline{\phi}_{\min}.$$

Suppose that there is a constant $\underline{C} > 0$ such that

$$\sigma_{\min}\left(I_{(d+K)p} - A\right) \geq \underline{C}\overline{\phi}_{\min} \tag{12}$$

In addition, there exists a positive constant $\overline{C} < \infty$ such that, for all positive integer j,

$$\sigma_{\max}\left(A^j\right) \leq \overline{C}\max\left\{\left|\lambda_{\max}\left(A^j\right)\right|, \left|\lambda_{\min}\left(A^j\right)\right|\right\}. \tag{13}$$

Note that Assumption 2-1 is the stability condition that one typically assumes for a stationary VAR process. One difference is that we allow for possible heterogeneity in the distribution of ε_t across time, so that our FAVAR process is not necessarily a strictly stationary process. Under Assumption 2-1, there exists a vector moving average representation for the FAVAR process. Note also that it is well known that $\det\left\{I_{(d+K)} - Az\right\} = \det\left\{I_{(d+K)} - A_1 z - \cdots - A_p z^p\right\}$, where A is the coefficient matrix of the companion form given in expression (5). See, for example, page 16 of Lütkepohl (2005). It follows that Assumption 2-1 is equivalent to the condition that $\det\left\{I_{(d+K)} - Az\right\} = 0$ implies that $|z| > 1$. In addition, Assumption 2-1 is equivalent to the assumption that all eigenvalues of A have modulus less than 1. Since the factor loading matrix Γ is an $N \times Kp$ matrix, where $N = N_1 + N_2$, the matrix $\Gamma'\Gamma$ will have order of magnitude equal to N if the factors are pervasive. Assumption 2-6 allows for possible violations of this conventional pervasiveness assumption, which will occur in our setup when $N_1/N \to 0$. Assumption 2-7 imposes a condition whereby the extreme singular values of the matrices A^j and $I_{(d+K)p} - A$ have bounds that depend on the extreme eigenvalues of these matrices. More primitive conditions for such a relationship between the singular values and the eigenvalues of a (not necessarily symmetric) matrix have been studied in the linear algebra literature. In Appendix B of this paper, we prove one such

result which extends a well-known result by Ruhe (1975). More specifically, we state and prove the following lemma:

Lemma 2-1: *Let A be an $n \times n$ square matrix with (ordered) singular values given by $\sigma_{(1)}(A) \geq \sigma_{(2)}(A) \geq \cdots \geq \sigma_{(n)}(A) \geq 0$. Suppose that A is diagonalizable, i.e., $A = S\Lambda S^{-1}$, where Λ is diagonal matrix whose diagonal elements are the eigenvalues of A. Let the modulus of these eigenvalues be ordered as follows: $\left|\lambda_{(1)}(A)\right| \geq \left|\lambda_{(2)}(A)\right| \geq \cdots \geq \left|\lambda_{(n)}(A)\right|$. Then, for $k \in \{1, ..., n\}$ and for any positive integer j, we have that:*

$$\chi(S)^{-1}\left|\lambda_{(k)}\left(A^j\right)\right| \leq \sigma_{(k)}\left(A^j\right) \leq \chi(S)\left|\lambda_{(k)}\left(A^j\right)\right|$$

where

$$\chi(S) = \sigma_{(1)}(S)\,\sigma_{(1)}\left(S^{-1}\right).^{6}$$

Note that in the special case where the matrices A and $I_{(d+K)p} - A$ are diagonalizable, the inequalities given in expressions (12) and (13) are a direct consequence of this lemma. On the other hand, Assumption 2-7 takes into account other situations where expressions (12) and (13) are valid even though the matrices A and $I_{(d+K)p} - A$ are not diagonalizable.

Assumptions 2-1, 2-2(a)-(c), and 2-7 together allow us to show in Lemma B-11 of Appendix B that the process $\{W_t\}$ generated by the FAVAR model given in expression (1) is a β-mixing process with β-mixing coefficient satisfying:

$$\beta_W(m) \leq a_1 \exp\{-a_2 m\},$$

for some positive constants a_1 and a_2, with

$$\beta_W(m) = \sup_t E\left[\sup\left\{\left|P\left(B|\mathcal{A}_{-\infty}^t\right) - P(B)\right| : B \in \mathcal{A}_{t+m}^{\infty}\right\}\right],$$

and with $\mathcal{A}_{-\infty}^t = \sigma(..., W_{t-2}, W_{t-1}, W_t)$ and $\mathcal{A}_{t+m}^{\infty} = \sigma(W_{t+m}, W_{t+m+1}, W_{t+m+2}, ...)$. Note that Assumption 2-2 (c) rules out situations such as that given in the famous counterexample presented by Andrews (1984) which shows that a first-order autoregression with errors having a discrete Bernoulli distribution is not α-mixing, even if it satisfies the stability condition. Conditions similar to Assumption 2-2(c) have also appeared in previous papers, such as Gorodetskii (1977) and Pham and Tran (1985), which seek to provide sufficient conditions for establishing the α or β mixing properties of linear time series processes.

Our variable selection procedure is based on a self-normalized statistic and makes use of some pathbreaking moderate deviation results for weakly dependent processes recently obtained by Chen,

Shao, Wu, and Xu (2016). An advantage of using a self-normalized statistic, as we will discuss a bit more following expression (18) below, is that it allows the range of the moderate deviation approximation to be wider relative to their non-self-normalized counterparts. To accommodate data dependence, we consider self-nomalized statistics that are constructed from observations which are first split into blocks in a manner similar to the kind of construction one would employ in implementing a block bootstrap or in proving a central limit theorem using the blocking technique. Two such statistics are proposed in this paper. The first of these statistics has the form of an ℓ_∞ norm and is given by:

$$\max_{1\leq \ell \leq d} |S_{i,\ell,T}| = \max_{1\leq \ell \leq d} \left| \frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}} \right|, \tag{14}$$

where

$$\overline{S}_{i,\ell,T} = \sum_{r=1}^{q} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \text{ and} \tag{15}$$

$$\overline{V}_{i,\ell,T} = \sum_{r=1}^{q} \left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} Z_{it} y_{\ell,t+1} \right]^2. \tag{16}$$

Here, Z_{it} denotes the i^{th} component of Z_t , $y_{\ell,t+1}$ denotes the ℓ^{th} component of Y_{t+1}, $\tau_1 = \lfloor T_0^{\alpha_1} \rfloor$, and $\tau_2 = \lfloor T_0^{\alpha_2} \rfloor$, where $1 > \alpha_1 \geq \alpha_2 > 0$, $\tau = \tau_1 + \tau_2$, $q = \lfloor T_0/\tau \rfloor$, and $T_0 = T - p + 1$. Note that the statistic given in expression (14) can be interpreted as the maximum of the (self-normalized) sample covariances between the i^{th} component of Z_t and the components of Y_{t+1}. Our second statistic has the form of a pseudo-L_1 norm and is given by:

$$\sum_{\ell=1}^{d} \varpi_\ell \, |S_{i,\ell,T}| = \sum_{\ell=1}^{d} \varpi_\ell \left| \frac{\overline{S}_{i,\ell,T}}{\sqrt{\overline{V}_{i,\ell,T}}} \right|,$$

where $\overline{S}_{i,\ell,T}$ and $\overline{V}_{i,\ell,T}$ are as defined in expressions (15) and (16) above and where $\{\varpi_\ell : \ell = 1, .., d\}$ denotes pre-specified weights, such that $\varpi_\ell \geq 0$, for every $\ell \in \{1, ..., d\}$, and $\sum_{\ell=1}^{d} \varpi_\ell = 1$. Both of these statistics employ a blocking scheme similar to that proposed in Chen, Shao, Wu, and Xu (2016), where, in order to keep the effects of dependence under control, the construction of these statistics is based only on observations in every other block. To see this, note that if we write out

the "numerator" term $\overline{S}_{i,\ell,T}$ in greater detail, we have that:

$$\begin{aligned}\overline{S}_{i,\ell,T} &= \sum_{t=p}^{\tau_1+p-1} Z_{it}y_{\ell,t+1} + \sum_{t=\tau+p}^{\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \\ &+ \sum_{t=2\tau+p}^{2\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} + \cdots + \sum_{t=(q-1)\tau+p}^{(q-1)\tau+\tau_1+p-1} Z_{it}y_{\ell,t+1} \end{aligned} \tag{17}$$

Comparing the first term and the second terms on the right-hand side of expression (17), we see that the observations $Z_{it}y_{\ell,t+1}$, for $t = \tau_1+p, ..., \tau+p-1$, have not been included in the construction of the sum. Similar observations hold when comparing the second and the third terms, and so on.

It should also be pointed out that although Chen, Shao, Wu, and Xu (2016) focus their analysis on problems of testing and inference for the mean of a scalar weakly dependent time series using self-normalized Student-type test statistics, our paper applies the self-normalization approach to a variable selection problem in a FAVAR setting. Namely, the problem which we study is more akin to a classification (or model selection) problem rather than a multiple hypothesis testing problem. In order to consistently estimate the factors up to an invertible matrix transformation, we develop a variable selection procedure whereby both the probability of a false positive and the probability of a false negative converge to zero as N_1, N_2, $T \to \infty$[7]. This is different from the typical multiple hypothesis testing approach whereby one tries to control the familywise error rate (or, alternatively, the false discovery rate), so that it is no greater than 0.05, say, but does not try to ensure that this probability goes to zero as the sample size grows.

To determine whether the i^{th} component of Z_t is a relevant variable for the purpose of factor estimation, we propose the following procedure. Define $i \in \widehat{H}^c$ to indicate that the procedure has classified Z_{it} to be a relevant variable for the purpose of factor estimation. Similarly, define $i \in \widehat{H}$ to indicate that the procedure has classified Z_{it} to be an irrelevant variable. Now, let $\mathbb{S}^+_{i,T}$ denote either the statistic $\max_{1\leq\ell\leq d} |S_{i,\ell,T}|$ or the statistic $\sum_{\ell=1}^{d} \varpi_\ell |S_{i,\ell,T}|$. Our variable selection procedure is based on the decision rule:

$$i \in \begin{cases} \widehat{H}^c & \text{if } \mathbb{S}^+_{i,T} \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \\ \widehat{H} & \text{if } \mathbb{S}^+_{i,T} < \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \end{cases}, \tag{18}$$

where $\Phi^{-1}(\cdot)$ denotes the quantile function or the inverse of the cumulative distribution function of the standard normal random variable, and where φ is a tuning parameter which may depend on

[7]Here, a false positive refers to mis-classifying a variable, Z_{it}, as a relevant variable for the purpose of factor estimation when its factor loading $\gamma_i' = 0$, whereas a false negative refers to the opposite case, where $\gamma_i' \neq 0$, but the variable Z_{it} is mistakenly classified as irrelevant.

N. Some conditions on φ will be given in Assumptions 2-11 and 2-11* below.

To understand why using the quantile function of the standard normal as the threshold function for our procedure is a natural choice, note first that, by a slight modification of the arguments given in the proof of Lemma B-17 in Appendix B of the Technical Appendix, we can show that, as $T \to \infty$

$$P\left(|S_{i,\ell,T}| \geq z\right) = 2\left[1 - \Phi\left(z\right)\right]\left(1 + o\left(1\right)\right), \tag{19}$$

which holds for all i and ℓ and for all z such that
$0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}/L\left(T\right), T^{\alpha_2/2}\right\}$, where $L\left(T\right)$ denotes a slowly varying function such that $L\left(T\right) \to \infty$ but $L\left(T\right)/T^{(1-\alpha_1)/6} \to 0$ as $T \to \infty$. In view of expression (19), we can interpret moderate deviation as providing an asymptotic approximation of the (two-sided) tail behavior of the self-normalized statistic, $S_{i,\ell,T}$, based on the tails of the standard normal distribution. An important advantage of using self-normalized statistics in this context is that the range for which this standard normal approximation is valid (i.e., the range $0 \leq z \leq c_0 \min\left\{T^{(1-\alpha_1)/6}/L\left(T\right), T^{\alpha_2/2}\right\}$) is wider for self-normalized statistics relative to their non-self-normalized counterparts. Now, suppose initially that we wish simply to control the probability of a Type I error for testing the null hypothesis $H_0 : \gamma_i = 0$ (i.e., the i^{th} variable does not load on the underlying factors) at some fixed significance level α. Then, expression (19) suggests that a natural way to do this is to set $z = \Phi^{-1}\left(1 - \alpha/2\right)$. This is because, given that the quantile function $\Phi^{-1}\left(\cdot\right)$ is, by definition, the inverse function of the cdf $\Phi\left(\cdot\right)$, we have that:

$$P\left(|S_{i,\ell,T}| \geq \Phi^{-1}\left(1 - \alpha/2\right)\right) = 2\left[1 - \Phi\left(\Phi^{-1}\left(1 - \alpha/2\right)\right)\right]\left(1 + o\left(1\right)\right) = \alpha\left(1 + o\left(1\right)\right),$$

so that the probability of a Type I error is controlled at the desired level α asymptotically. Note also that an advantage of moderate deviation theory is that it gives a characterization of the relative approximation error, as opposed to the absolute approximation error. As a result, the approximation given is useful and meaningful even when α is very small, which is of importance to us since we are interested in situations where we might want to let α go to zero, as sample size approaches infinity.

The above example provides intuition concerning the form of the threshold function that we have specified. The variable selection problem that we actually consider is more complicated however, since we need to control the probability of a Type I error (or of a false positive) not just for a single test involving the i^{th} variable but for a multiple hypothesis testing scenario involving the loading coefficient vectors for all variables Z_{it} (for $i = 1, ..., N$). Moreover, as noted previously, we want also to design a procedure where the probability of a false positive will go asymptotically to

zero as well. We show in Theorem 1 below that these objectives can all be accomplished using the threshold function specified in expression (18).[8]

Indeed, under appropriate conditions, the variable selection procedure described above can be shown to be completely consistent, in the sense that both the probability of a false positive, i.e. $P\left(i \in \widehat{H}^c | i \in H\right)$, and the probability of a false negative, i.e., $P\left(i \in \widehat{H} | i \in H^c\right)$, approach zero as $N, T \to \infty$. To show this result, we must first state a number of additional assumptions.

Assumption 2-8: There exists a positive constant, $\underline{c}$, such that for T sufficiently large:

$$\min_{1\leq \ell \leq d} \min_{i \in H} \min_{r \in \{1,...,q\}} E\left\{ \left[\frac{1}{\sqrt{\tau_1}} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} y_{\ell,t+1} u_{it} \right]^2 \right\} \geq \underline{c},$$

where, as defined earlier,

$$\tau_1 = \lfloor T_0^{\alpha_1} \rfloor, \ \tau_2 = \lfloor T_0^{\alpha_2} \rfloor \text{ for } 1 > \alpha_1 \geq \alpha_2 > 0 \text{ and } q = \left\lfloor \frac{T_0}{\tau_1 + \tau_2} \right\rfloor,$$

and $T_0 = T - p + 1$.

Assumption 2-9: Let $i \in H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}$. Suppose that there exists a positive constant, $\underline{c}$, such that, for all N_1, N_2,and T sufficiently large:

$$\begin{aligned} & \min_{1\leq \ell \leq d} \min_{i \in H^c} \left| \frac{\mu_{i,\ell,T}}{q\tau_1} \right| \\ = & \min_{1\leq \ell \leq d} \min_{i \in H^c} \left| \frac{1}{q} \sum_{r=1}^{q} \frac{1}{\tau_1} \sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} \gamma_i' \left\{ E\left[\underline{F}_t\right] \mu_{Y,\ell} + E\left[\underline{F}_t \underline{Y}_t'\right] \alpha_{YY,\ell} + E\left[\underline{F}_t \underline{F}_t'\right] \alpha_{YF,\ell} \right\} \right| \\ \geq & \ \underline{c} > 0, \end{aligned}$$

where $\mu_{Y,\ell} = e_{\ell,d}' \mu_Y$, $\alpha_{YY,\ell} = A_{YY}' e_{\ell,d}$, and $\alpha_{YF,\ell} = A_{YF}' e_{\ell,d}$. Here, $e_{\ell,d}$ is a $d \times 1$ elementary vector whose ℓ^{th} component is 1 and all other components are 0.

Assumption 2-10: Suppose that, as N_1, N_2, and $T \to \infty$, the following rate conditions hold: $\frac{\sqrt{\ln N}}{T^{\min\left\{\frac{1-\alpha_1}{6}, \frac{\alpha_2}{2}\right\}}} \to 0$, where (a) $1 > \alpha_1 \geq \alpha_2 > 0$ and $N = N_1 + N_2$, and (b) $N_1/T^{3\alpha_1} \to 0$ where $1 > \alpha_1 > 0$.

Assumption 2-11: Let φ satisfy the following two conditions: (a) $\varphi \to 0$ as $N_1, N_2 \to \infty$, and (b)

[8] The threshold function used here is reminiscent of the one employed in a celebrated paper by Belloni, Chen, Chernozhukov, and Hansen (2012). More specifically, Belloni, Chen, Chernozhukov, and Hansen (2012) use a similar threshold function to help set the penalty level for Lasso estimation of the first-stage equation of an IV regression model assuming *i.n.i.d.* data. In spite of the similarity in the form of the threshold function, the problem studied in that paper is very different from the one which we analyze here. In consequence, the conditions we specify for setting the tuning parameter φ will also be quite different from what they recommend in their paper.

there exists some constant $a > 0$, such that $\varphi \geq 1/N^a$, for all N_1, N_2 sufficiently large. Assumption 2-8 rules out certain degenerate situations where as $T \to \infty$

$$E\left\{\left[\frac{1}{\sqrt{\tau_1}}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1} y_{\ell,t+1}u_{it}\right]^2\right\} \to 0$$

for some $1 \leq \ell \leq d$, $i \in H$, and $r \in \{1, .., q\}$, since a moderate deviation result and, in fact, a central limit theorem, would not hold in general if such degeneracies were to occur. A similar condition is also assumed in Chen, Shao, Wu, and Xu (2016). See condition (4.2) on page 1600 of that paper.

To give an intuitive interpretation for Assumption 2-9, note that the term

$$\frac{\mu_{i,\ell,T}}{q\tau_1} = \frac{1}{q}\sum_{r=1}^{q}\frac{1}{\tau_1}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}\gamma_i'\left\{E\left[\underline{F}_t\right]\mu_{Y,\ell} + E\left[\underline{F}_t\underline{Y}_t'\right]\alpha_{YY,\ell} + E\left[\underline{F}_t\underline{F}_t'\right]\alpha_{YF,\ell}\right\}$$

is, in fact, a noncentrality parameter for the variable-selection/multiple-hypothesis-testing problem considered here. This condition allows us to differentiate between the null hypothesis, $i \in H$ (where $\gamma_i = 0$) from the alternative hypothesis $i \in H^c$ (where $\gamma_i \neq 0$) so that, under this condition, it will be possible to design procedures, such as the one proposed here, which will have asymptotic power. To see this more clearly, note that if $i \in H$, then it is clear that:

$$\frac{\mu_{i,\ell,T}}{q\tau_1} = \frac{1}{q}\sum_{r=1}^{q}\frac{1}{\tau_1}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}\gamma_i'\left\{E\left[\underline{F}_t\right]\mu_{Y,\ell} + E\left[\underline{F}_t\underline{Y}_t'\right]\alpha_{YY,\ell} + E\left[\underline{F}_t\underline{F}_t'\right]\alpha_{YF,\ell}\right\} = 0,$$

given that $\gamma_i = 0$ in this case. On the other hand, under the alternative hypothesis where $i \in H^c$, we will have $\gamma_i \neq 0$ so that $\mu_{i,\ell,T}/\left(q\tau_1\right) \neq 0$ under Assumption 2-9. Now, we believe Assumption 2-9 is a fairly mild condition to be placed on a FAVAR since, given the interconnectedness of a FAVAR, it is unlikely to have a situation where

$$\frac{\mu_{i,\ell,T}}{q\tau_1} = \frac{1}{q}\sum_{r=1}^{q}\frac{1}{\tau_1}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}\gamma_i'\left\{E\left[\underline{F}_t\right]\mu_{Y,\ell} + E\left[\underline{F}_t\underline{Y}_t'\right]\alpha_{YY,\ell} + E\left[\underline{F}_t\underline{F}_t'\right]\alpha_{YF,\ell}\right\} = 0$$

in the case where $\gamma_i \neq 0$. It should be noted, of course, that Assumption 2-9 does rule out certain specialized situations, such as the case when $\mu_{Y,\ell} = 0$, $\alpha_{YY,\ell} = 0$, and $\alpha_{YF,\ell} = 0$, for some $\ell \in \{1, ..., d\}$. However, we do not consider such cases to be of much practical interest since, for example, if $\mu_{Y,\ell} = 0$, $\alpha_{YY,\ell} = 0$, and $\alpha_{YF,\ell} = 0$ for some ℓ then expression (2) above implies that

the ℓ^{th} component of Y_{t+1} will have the representation

$$\begin{aligned} y_{\ell,t+1} &= \mu_{Y,\ell} + \underline{Y}_t'\alpha_{YY,\ell} + \underline{F}_t'\alpha_{YF,\ell} + \varepsilon_{\ell,t+1}^Y \\ &= \varepsilon_{\ell,t+1}^Y, \end{aligned}$$

so that, in this case, $y_{\ell,t+1}$ depends neither on $\underline{Y}_t = \left(Y_t', Y_{t-1}', ..., Y_{t-p+1}'\right)'$ nor on $\underline{F}_t = \left(F_t', F_{t-1}', ..., F_{t-p+1}'\right)$. This is, of course, an unrealistic model for $y_{\ell,t+1}$ since it would not even be dependent. Hence, we do not expect Assumption 2-9 to be violated except in highly degenerate situations such as the one just described.

The following two theorems give our main theoretical results on the variable selection procedure described above.

Theorem 1: *Let $H = \{k \in \{1, ..., N\} : \gamma_k = 0\}$. Suppose that Assumptions 2-1, 2-2(a)-(c), 2-3(a)-(c) 2-4, 2-5, 2-7, 2-8, 2-10 (a) and 2-11 hold. Let $\Phi^{-1}(\cdot)$ denote the inverse of the cumulative distribution function of the standard normal random variable, or, alternatively, the quantile function of the standard normal distribution. Then, the following statements are true:*

(a) *Let $\{\varpi_\ell : \ell = 1, .., d\}$ be pre-specified weights, such that $\varpi_\ell \geq 0$, for every $\ell \in \{1, ..., d\}$ and $\sum_{\ell=1}^d \varpi_\ell = 1$, then:*

$$P\left(\max_{i\in H}\sum_{\ell=1}^d \varpi_\ell \left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) = O\left(\frac{N_2\varphi}{N}\right) = o(1),$$

where $N = N_1 + N_2$.

(b)

$$P\left(\max_{i\in H}\max_{1\leq\ell\leq d}\left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) = O\left(\frac{N_2\varphi}{N}\right) = o(1).$$

Theorem 2: *Let $H^c = \{k \in \{1, ..., N\} : \gamma_k \neq 0\}$. Suppose that Assumptions 2-1, 2-2(a)-(c), 2-3(a)-(c), 2-5, 2-7, 2-9, 2-10, and 2-11 hold. Then, the following statements are true.*

(a) *Let $\{\varpi_\ell : \ell = 1, .., d\}$ be pre-specified weights, such that $\varpi_\ell \geq 0$, for every $\ell \in \{1, ..., d\}$ and $\sum_{\ell=1}^d \varpi_\ell = 1$, then:*

$$P\left(\min_{i\in H^c}\sum_{\ell=1}^d \varpi_\ell \left|S_{i,\ell,T}\right| \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)\right) \to 1.$$

(b)

$$P\left(\min_{i\in H^c}\max_{1\leq\ell\leq d}|S_{i,\ell,T}|\geq\Phi^{-1}\left(1-\frac{\varphi}{2N}\right)\right)\to 1.$$

Theorem 1 shows that, under our procedure, the probability of a false positive, i.e., the probability that $i\in\widehat{H}^c$, even though $\gamma_i=0$, approaches zero, as $N,T\to\infty$. Theorem 2 shows that the probability of a false negative, i.e., the probability that $i\in\widehat{H}$ even though $\gamma_i\neq 0$, also approaches zero, as $N,T\to\infty$. Together, these two theorems show that our variable selection procedure is (completely) consistent in the sense that the probability of committing a misclassification error vanishes as $N,T\to\infty$.

Remark 2.1:

It should be noted that a special case of our FAVAR model which is of particular interest is the case where $d=1$, i.e., the case where the Y variable is univariate. In this case, equation (2) reduces to

$$y_{t+1}=\mu_Y+a_{YY,1}y_t+\cdots+a_{YY,p}y_{t-p+1}+a'_{YF}\underline{F}_t+\varepsilon^Y_{t+1} \tag{20}$$

where μ_Y is now a 1×1 intercept parameter; $a_{YY,1}...,a_{YY,p}$ are the p autoregressive parameters; a'_{YF} is a $1\times Kp$ coefficient vector; and ε^Y_{t+1} is now a 1×1 error term. Expression (20), thus, yields a factor augmented autoregressive model which is commonly used to forecast economic time series. Moreover, when $d=1$, it is easy to see that our two statistics, $\max_{1\leq\ell\leq d}|S_{i,\ell,T}|$ and $\sum_{\ell=1}^{d}\varpi_\ell|S_{i,\ell,T}|$, reduce to the same one since

$$\max_{1\leq\ell\leq d}|S_{i,\ell,T}|=\max_{1\leq\ell\leq 1}|S_{i,\ell,T}|=S_{i,1,T}=\sum_{\ell=1}^{1}\varpi_\ell|S_{i,\ell,T}|=\sum_{\ell=1}^{d}\varpi_\ell|S_{i,\ell,T}|$$

given that $1=\sum_{\ell=1}^{d}\varpi_\ell=\varpi_1$ in this case. Hence, for the $d=1$ case, we can remove the subscript 1 and write

$$S_{i,1,T}=S_{i,T}=\frac{\overline{S}_{i,T}}{\sqrt{\overline{V}_{i,T}}} \tag{21}$$

where

$$\overline{S}_{i,T}=\sum_{r=1}^{q}\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}Z_{it}y_{t+1}\text{ and }\overline{V}_{i,\ell,T}=\sum_{r=1}^{q}\left[\sum_{t=(r-1)\tau+p}^{(r-1)\tau+\tau_1+p-1}Z_{it}y_{t+1}\right]^2 \tag{22}$$

so that our statistic can be viewed as a self-normalized sample covariance constructed using blocked sums.

Remark 2.2:

It is also worth stressing at this point that although we interpret our variable selection procedure

primarily as a procedure which assesses the relevance of a variable Z_{it}, it is clear from expressions (21) and (22) above that our procedure also contains information about the predictive content of the variable Z_{it} for y_{t+1}. There is a close association between the relevance and the predictive content of a variable in the context of a FAVAR model. This is true because if $\gamma_i = 0$, i.e., if Z_{it} does not load on any of the underlying factors, so that Z_{it} is an irrelevant variable; then Z_{it} will also not have predictive content for y_{t+1} because, within a FAVAR system, any possible correlation between Z_{it} and y_{t+1} only works its way through indirectly via the factors. Hence, if Z_{it} is not correlated with any of the factors, then it will not have correlation with y_{t+1}. However, the reason why we choose to interpret our procedure as one which primarily assesses the relevance of the variables Z_{it} (for $i = 1, ..., N$) is because a FAVAR system is quite complex, and one can find examples where a Z_{it} variable is specified to not have any predictive content for y_{t+1}, but a variable selection procedure based on the score statistic nevertheless rejects the null hypothesis with probability one as sample sizes approach infinity. To see how this could be the case, consider the following example.

Example 1: Consider a two-factor FAVAR model of the form:

$$\begin{aligned} y_{t+1} &= a_{YY} y_t + \alpha_{YF,1} f_{1,t} + \varepsilon^Y_{t+1} \\ f_{1,t+1} &= a_{FY,1} y_t + a_{FF,11} f_{1,t} + a_{FF,12} f_{2,t} + \varepsilon^F_{1,t+1} \\ f_{2,t+1} &= a_{FY,2} y_t + a_{FF,21} f_{1,t} + a_{FF,22} f_{2,t} + \varepsilon^F_{2,t+1}, \end{aligned} \tag{23}$$

with the factor equation given by

$$Z_t = \Gamma F_t + u_t, \tag{24}$$

where $F_t = \begin{pmatrix} f_{1,t} & f_{2,t} \end{pmatrix}'$ and where $\alpha_{YF,1} \neq 0$. Note that, under the specification given by expressions (23) and (24), the factor $f_{2,t}$ has no predictive content for the target variable of interest y_{t+1}, whereas the factor $f_{1,t}$ does have predictive content. Now, write the companion form:

$$W_{t+1} = AW_t + \varepsilon_{t+1},$$

where

$$W_t = \begin{pmatrix} y_t \\ f_{1,t} \\ f_{2,t} \end{pmatrix}, \varepsilon_t = \begin{pmatrix} \varepsilon^Y_t \\ \varepsilon^F_{1,t} \\ \varepsilon^F_{2,t} \end{pmatrix}, \text{ and } A = \begin{pmatrix} a_{YY} & \alpha_{YF,1} & 0 \\ a_{FY,1} & a_{FF,11} & a_{FF,12} \\ a_{FY,2} & a_{FF,21} & a_{FF,22} \end{pmatrix},$$

and where we defne $\Sigma_\varepsilon = E\left[\varepsilon_t \varepsilon_t'\right]$ and assume that Σ_ε is positive definite. Here, under Assumption

2-1, we have the vector moving-average representation:

$$W_{t+1} = \sum_{j=0}^{\infty} A^j \varepsilon_{t+1},$$

It follows that the components of W_{t+1} have the univariate MA representations:

$$y_{t+1} = \sum_{j=0}^{\infty} e_1' A^j \varepsilon_{t+1-j},\ f_{1,t} = \sum_{k=0}^{\infty} e_2' A^k \varepsilon_{t-k}, \text{ and } f_{2,t} = \sum_{k=0}^{\infty} e_3' A^k \varepsilon_{t-k},$$

with $e_1 = \begin{pmatrix} 1 & 0 & 0 \end{pmatrix}'$, $e_2 = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}'$, and $e_3 = \begin{pmatrix} 0 & 0 & 1 \end{pmatrix}'$. Let Z_{it} and Z_{jt} be, respectively, the i^{th} and the j^{th} components of Z_t (with $i \neq j$). Suppose that Z_{it} loads only on the second factor but not the first, so that $\gamma_i' = \begin{pmatrix} 0 & \gamma_{i2} \end{pmatrix}$, where $\gamma_{i2} \neq 0$; and suppose that Z_{jt} loads only on the first factor but not the second, so that $\gamma_j' = \begin{pmatrix} \gamma_{j1} & 0 \end{pmatrix}$, where $\gamma_{j1} \neq 0$. Hence, both Z_{it} and Z_{jt} are relevant variables for factor estimation, but Z_{jt} has predictive content for y_{t+1} whereas Z_{it} does not. Consider the score statistics associated with Z_{it} and Z_{jt} :

$$\begin{aligned} S_i &= \sum_{t=1}^{T-1} Z_{it} y_{t+1} = \sum_{t=1}^{T-1} \left(\gamma_i' F_t + u_{it}\right) y_{t+1} = \sum_{t=1}^{T-1} \gamma_{i2} f_{2,t} y_{t+1} + \sum_{t=1}^{T-1} u_{it} y_{t+1} \text{ and} \\ S_j &= \sum_{t=1}^{T-1} Z_{jt} y_{t+1} = \sum_{t=1}^{T-1} \left(\gamma_j' F_t + u_{jt}\right) y_{t+1} = \sum_{t=1}^{T-1} \gamma_{j1} f_{1,t} y_{t+1} + \sum_{t=1}^{T-1} u_{jt} y_{t+1}. \end{aligned}$$

Now, suppose that $\sigma_{32} \neq 0$; (where σ_{32} denotes the $(3,2)^{th}$ element of the error covariance matrix Σ_ε); then, given that $\gamma_{j2} \neq 0$ and $\alpha_{YF,1} \neq 0$, the expected value of S_i will not, in general, be properly centered at zero. That is,

$$\begin{aligned} E[S_i] &= \sum_{t=1}^{T-1} E[Z_{it} y_{t+1}] = \gamma_{i2} \sum_{t=1}^{T-1} E[f_{2,t} y_{t+1}] + \sum_{t=1}^{T-1} E[u_{it} y_{t+1}] \\ &= \gamma_{i2} \sum_{t=1}^{T-1} \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} e_3' A^k E\left[\varepsilon_{t-k} \varepsilon_{t+1-\ell}'\right] \left(A'\right)^\ell e_1 = \gamma_{i2} \sum_{t=1}^{T-1} \sum_{k=0}^{\infty} e_3' A^k \Sigma_\varepsilon \left(A'\right)^{k+1} e_1 \\ &= \gamma_{i2} (T-1) \left[\sigma_{31} a_{YY} + \sigma_{32} \alpha_{YF,1}\right] + \gamma_{i2} \sum_{t=1}^{T-1} \sum_{k=1}^{\infty} e_3' A^k \Sigma_\varepsilon \left(A'\right)^{k+1} e_1 \\ &\neq 0, \end{aligned}$$

except in very specialized cases.[9] Moreover, given that $\gamma_{j1} \neq 0$, $\sigma_{22} > 0$, and $\alpha_{YF,1} \neq 0$; the

[9] The reason why we say that in this case $E[S_i] \neq 0$, except in very specialized cases, is because although given

expected value of S_j will also not, in general, be properly centered at zero. That is,

$$\begin{aligned}
E[S_j] &= \sum_{t=1}^{T-1} E[Z_{jt} y_{t+1}] \\
&= \gamma_{j1} \sum_{t=1}^{T-1} E[f_{1,t} y_{t+1}] + \sum_{t=1}^{T-1} E[u_{jt} y_{t+1}] \\
&= \gamma_{j1} \sum_{t=1}^{T-1} \sum_{k=0}^{\infty} \sum_{\ell=0}^{\infty} e_2' A^k E\left[\varepsilon_{t-k} \varepsilon_{t+1-\ell}'\right] \left(A'\right)^{\ell} e_1 \\
&= \gamma_{j1} \sum_{t=1}^{T-1} \sum_{k=0}^{\infty} e_2' A^k \Sigma_{\varepsilon} \left(A'\right)^{k+1} e_1 \\
&= \gamma_{j1} (T-1) [\sigma_{21} a_{YY} + \sigma_{22} \alpha_{YF,1}] + \gamma_{j1} \sum_{t=1}^{T-1} \sum_{k=1}^{\infty} e_2' A^k \Sigma_{\varepsilon} \left(A'\right)^{k+1} e_1 \\
&\neq 0,
\end{aligned}$$

except in very specialized cases. Hence, both statistics, when appropriately normalized, will diverge with probability approaching one, as $T \to \infty$. This makes the right inference about the relevance of both of these variables, since the divergence of these statistics implies that the null hypothesis $H_0 : \gamma_i = 0$ (i.e., Z_{it} is irrelevant) as well as the null hypothesis $H_0 : \gamma_j = 0$ (i.e., Z_{jt} is irrelevant) will both be rejected with probability approaching one. However, if we were to interpret these statistics as providing inference about the predictive content of the variables Z_{it} and Z_{jt}; then, we would have made the wrong inference about Z_{it}, since it loads only on $f_{2,t}$ which is not helpful in predicting y_{t+1}.

On the other hand, consider the alternative scenario where $\gamma_{i2} = 0$ and $\gamma_{j1} = 0$ so that $\gamma_i' = \begin{pmatrix} 0 & \gamma_{i2} \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$ and $\gamma_j' = \begin{pmatrix} \gamma_{j1} & 0 \end{pmatrix} = \begin{pmatrix} 0 & 0 \end{pmatrix}$, and, thus, both Z_{it} and Z_{jt} are

that $\gamma_{i2} \neq 0$, $\sigma_{32} \neq 0$, and $\alpha_{YF,1} \neq 0$; it is clear that the term $\gamma_{i2} (T-1) \sigma_{32} \alpha_{YF,1} \neq 0$; one may nevertheless argue that even in this case it is possible to have $E[S_i] = 0$ if it turns out that

$$\gamma_{i2} (T-1) \sigma_{32} \alpha_{YF,1} = -\gamma_{i2} (T-1) \sigma_{31} a_{YY} - \gamma_{i2} \sum_{t=1}^{T-1} \sum_{k=1}^{\infty} e_3' A^k \Sigma_{\varepsilon} \left(A'\right)^{k+1} e_1.$$

However, note that for the above identity to hold, the elements of A and Σ_{ε} including σ_{31} and a_{YY} must take on very specific values so that in general the above identity is not likely to hold, in which case we would have $E[S_i] \neq 0$.

now irrelevant variables. Then, under this alternative scenario, we would have

$$\begin{aligned} E[S_i] &= \gamma_{i2} \sum_{t=1}^{T-1} \sum_{j=1}^{\infty} e_3' A^{j-1} \Sigma_\varepsilon \left(A'\right)^j e_1 = 0 \\ E[S_j] &= \gamma_{j1} \sum_{t=1}^{T-1} \sum_{j=1}^{\infty} e_2' A^{j-1} \Sigma_\varepsilon \left(A'\right)^j e_1 = 0 \end{aligned}$$

so that both statistics are now properly centered at zero, and neither will diverge, when appropriately normalized, as $T \to \infty$. It follows that, under this alternative scenario, given an appropriate threshold or critical value, we will also make the correct inference asymptotically about the fact that both Z_{it} and Z_{jt} are irrelevant variables in this case. Hence, under both scenarios, we make the right inference asymptotically about the relevance of a variable; however, under the first scenario, we do not make the right inference about predictive content. For ease of presentation, we have given an example based on a simple score statistic whose construction does not involve a blocking scheme or self-normalization. The same story holds, however, for the more complicated score statistics discussed in this paper.

The above example shows that interpreting a score-statistic-based variable selection procedure as a procedure which selects variables based on predictive content as opposed to relevance leads to a situation where we cannot, in general, interpret these type of procedures as being completely consistent. However, leaving particular examples like this one aside, we do believe that it is often useful to interpret our variable selection procedure as being helpful both for assessing the predictive content and for assessing the relevance of a variable Z_{it} in a factor augmented forecasting context, since the two objectives are closely related. In addition, it should be noted that being able to correctly identify all of the relevant variables and use only the relevant variables to estimate the factors can itself be helpful from the perspective of forecasting. This is because a forecasting equation such as the one given in expression (20) above depends on the unobserved latent factors which must be estimated. In either the case where some irrelevant variables are employed in the estimation process or if some relevant variables are not employed, the quality of the factor estimates could be reduced; which, in turn, can adversely affect the quality of point forecasts.

Remark 2.3:

Note also that a valuable by-product of our variable selection procedure is that it provides us with an estimate $\widehat{N}_1$ of the unobserved quantity N_1, where the latter, in light of Assumption 2-6, can also be interpreted as giving the order of magnitude of $\Gamma'\Gamma$ and is, thus, a measure of the overall pervasiveness of the factors in a given application. As mentioned previously, since N_1 is itself not directly observable, having a consistent estimator $\widehat{N}_1$ provides practitioners with a useful

diagnostic statistic which can help them assess the overall pervasiveness of the factors in empirical applications. As we show in part (a) of Lemma C-15 in Appendix C of the Technical Appendix, $\widehat{N}_1$ is a consistent estimator of N_1, in the sense that $\widehat{N}_1/N_1 \xrightarrow{p} 1$. As can be seen from the proof of Lemma C-15(a), this consistency result will not be possible in general if we do not have a completely consistent variable selection procedure where the probabilities of both Type I error and Type II error vanish asymptotically. To give an intuitive example for why this is the case, consider the following.

Example 2: Suppose that $N_1 = \lceil (1-\alpha) N \rceil$ and $N_2 = \lfloor \alpha N \rfloor$ for some fixed α such that $0 < \alpha < 1$, and suppose that we use a variable selection procedure which, even in large sample, results in a 5% Type I error but no Type II error so that $\widehat{N}_1 = (N_1 + 0.05 N_2)[1 + o_p(1)]$. Then, it is easy to see that

$$\begin{aligned}
\frac{\widehat{N}_1}{N_1} - 1 &= \frac{\widehat{N}_1 - N_1}{N_1} \\
&= \frac{(N_1 + 0.05 N_2)[1 + o_p(1)] - N_1}{N_1} \\
&= \frac{0.05 N_2 [1 + o_p(1)]}{N_1} + o_p(1) \\
&= \frac{0.05 \lfloor \alpha N \rfloor}{\lceil (1-\alpha) N \rceil} [1 + o_p(1)] + o_p(1) \\
&= \frac{0.05 \lfloor \alpha N \rfloor}{\lceil (1-\alpha) N \rceil} + o_p(1) \\
&\neq o_p(1)
\end{aligned}$$

so that, under this scenario, $\widehat{N}_1$ is not a consistent estimator of N_1. Now, one may argue at this point that if we instead assume that $N_2/N_1 \to 0$ (which corresponds to the case where $\alpha \to 0$), then we will still have a consistency result such that $\widehat{N}_1/N_1 \xrightarrow{p} 1$, even when we use a variable selection procedure where the probability of a Type I error does not vanish asymptotically. However, note that if a dataset is well approximated by the rate condition $N_2/N_1 \to 0$, then the forecast results based on our procedure should be very similar to the forecast results obtained under the conventional PCA procedure, given that in this situation very few variables will be excluded from factor estimation under our variable selection procedure. This does not seem to be in accord with our empirical results reported in Section 5, where we find that using FRED-MD data delivers forecast results based on our procedure that are better than the forecast results of the conventional PCA procedure in a vast majority of the cases considered. Moreover, our estimates of N_1 and N_2 using the FRED-MD dataset are not indicative of a situation where N_2 is negligible relative to N_1.

Alternatively, suppose instead that we use a variable selection procedure which results in a 5%

Type II error but no Type I error in large sample, so that $\widehat{N}_1 = 0.95N_1 [1 + o_p(1)]$. For this case, we have

$$\begin{aligned}
\frac{\widehat{N}_1}{N_1} - 1 &= \frac{0.95N_1 [1 + o_p(1)] - N_1}{N_1} \\
&= -\frac{0.05N_1}{N_1} + o_p(1) \\
&= -0.05 + o_p(1) \\
&\neq o_p(1)
\end{aligned}$$

Together, the two cases described above show that, to obtain a consistent estimate of N_1, we need in general to apply a completely consistent variable selection method.

In their paper, Bai and Ng (2023) provide a rate condition for consistent factor estimation (i.e., Assumption A4 in their paper) which can be restated in our setup as the assumption that $N/(TN_1) \to 0$. Now, this rate condition can provide a very useful guide for applied researchers wishing to assess the overall pervasiveness of the factors in a particular empirical problem of interest to them if a consistent estimator can be developed for the unobserved quantity N_1, and this is exactly what our procedure supplies. Viewed from this perspective, what we are proposing here builds on the work of Bai and Ng (2023), as our procedure helps to highlight the importance of the rate condition they have introduced and provides additional information that will be helpful to empirical researchers in assessing the degree of pervasiveness of the underlying factors in a particular empirical application.

Remark 2.4:

Additionally, note that knowledge of the number of factors is not needed to implement our variable selection procedure. Hence, in the case where the number of factors needs to be determined empirically, an applied researcher could first use our procedure to properly select the relevant variables and then apply an information criterion such as that proposed in Bai and Ng (2002) to estimate the number of factors.

3 Consistent Estimation of the h-Step Ahead Predictor Based on the FAVAR Model

In this section, we provide our main theoretical results on factor estimation and also on the estimation of the h-step ahead predictor within the FAVAR framework. This includes deriving an explicit representation of the h-step ahead forecasting equation implied by the FAVAR model. The totality of our results, as provided in this section and in the previous section of the paper, gives a complete

description of our proposed methodology for constructing forecasts within a FAVAR framework. In particular, our results provide explicit formulae that allow empirical researchers to easily implement procedures for variable selection for the purpose of factor estimation, use the selected variables to construct estimators of the factors, and finally to estimate the h-step ahead predictor.

To obtain the results of this section, we need first to impose a further rate condition on the tuning parameter, φ (see part (c) of Assumption 2-11* below).

Assumption 2-11*: Let φ satisfy the following three conditions: (a) $\varphi \to 0$ as $N_1, N_2 \to \infty$; (b) there exists some constant $a > 0$, such that $\varphi \geq 1/N^a$, for all N_1, N_2 sufficiently large; and (c)

$$\max\left\{\frac{N^{\frac{2}{7}}\varphi^{\frac{5}{7}}}{N_1}, \frac{N^{\frac{1}{3}}\varphi}{N_1 T}\right\} \to 0 \text{ as } N_1, N_2, T \to \infty.$$

Note that the rate condition given in part (c) of Assumption 2-11* depends on N_1. However, if we choose φ so that $\varphi N^{\frac{2}{5}} = O(1)$, then

$$\frac{N^{\frac{2}{7}}\varphi^{\frac{5}{7}}}{N_1} = O\left(\frac{1}{N_1}\right) = o(1) \text{ and } \frac{N^{\frac{1}{3}}\varphi}{N_1 T} = O\left(\frac{1}{N_1 N^{\frac{1}{15}} T}\right) = o\left(\frac{1}{N_1}\right).$$

Hence, with this choice of φ, Assumption 2-11* part (c) will be satisfied as long as $N_1 \to \infty$, and there is no need to impose any further condition on the rate at which N_1 grows. Requiring that $N_1 \to \infty$ is a minimal condition, since if $N_1 \nrightarrow \infty$; then consistent factor estimation, even up to an invertible matrix transformation, is impossible. Moreover, Monte Carlo results reported in Section 5 of this paper show that our variable selection procedure performs very well in finite samples, under the tuning parameter choice $\varphi = N^{-\frac{2}{5}}$, both in terms of controlling the probability of a false positive (or Type I) error and in terms of controlling the probability of a false negative (or Type II) error.

Next, consider the post-variable-selection principal component estimator of $\underline{F}_t = \left(F_t', F_{t-1}', ..., F_{t-p+1}'\right)$:

$$\widehat{\underline{F}}_t = \frac{\widehat{\Gamma}' Z_{t,N}\left(\widehat{H^c}\right)}{\widehat{N}_1}, \tag{25}$$

where

$$Z_{t,N}\left(\widehat{H^c}\right) = \left[\begin{array}{cccc} Z_{1,t}\mathbb{I}\left\{1 \in \widehat{H^c}\right\} & Z_{2,t}\mathbb{I}\left\{2 \in \widehat{H^c}\right\} & \cdots & Z_{N,t}\mathbb{I}\left\{N \in \widehat{H^c}\right\} \end{array}\right]',$$

with

$$\mathbb{I}\left\{i \in \widehat{H^c}\right\} = \left\{\begin{array}{ll} 1 & \text{if } i \in \widehat{H^c}, \text{ i.e., if } \mathbb{S}_{i,T}^{+} \geq \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \\ 0 & \text{if } i \in \widehat{H}, \text{ i.e., if } \mathbb{S}_{i,T}^{+} < \Phi^{-1}\left(1 - \frac{\varphi}{2N}\right) \end{array}\right. ,$$

and where $\widehat{N}_1 = \#\left(\widehat{H^c}\right)$, i.e., the cardinality of the set $\widehat{H^c}$. Here, $\widehat{\Gamma}$ denotes the principal component estimator of the loading matrix Γ constructed from taking $\sqrt{\widehat{N}_1}$ times the eigenvectors of the post-variable-selection sample covariance matrix $\widehat{\Sigma}\left(\widehat{H^c}\right)$ associated with the Kp largest eigenvalues of this matrix, where, in this case, $\widehat{\Sigma}\left(\widehat{H^c}\right) = \frac{Z\left(\widehat{H^c}\right)' Z\left(\widehat{H^c}\right)}{\widehat{N}_1 T_0} = \frac{1}{\widehat{N}_1 T_0}\sum_{t=p}^{T} Z_{t,N}\left(\widehat{H^c}\right) Z_{t,N}\left(\widehat{H^c}\right)'$, with $T_0 = T - p + 1$. Our next result shows that the estimator given in expression (25) consistently estimates the unobserved factors $\underline{F}_t$,up to an invertible $Kp \times Kp$ matrix transformation.

Theorem 3: *Suppose that Assumptions 2-1, 2-2, 2-3, 2-4, 2-5, 2-6, 2-7, 2-8, 2-9, and 2-10 hold. Let $\widehat{\underline{F}}_t$ be as defined in expression (25). Assume further that the specification of the tuning parameter, φ, in the decision rule (18) satisfies Assumption 2-11*. Then,*

$$\left\|\widehat{\underline{F}}_t - Q'\underline{F}_t\right\|_2 = o_p(1), \text{ for all fixed } t,$$

where

$$Q = \left(\frac{\Gamma'\Gamma}{N_1}\right)^{\frac{1}{2}} \Xi \widehat{V},$$

and where $\widehat{V}$ is the $Kp \times Kp$ orthogonal matrix given in Lemma C-14 part (c), and Ξ is a $Kp \times Kp$ orthogonal matrix whose columns are the eigenvectors of the matrix

$$M^*_{FF} = \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2} M_{FF} \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2} = \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2} \frac{1}{T_0}\sum_{t=p}^{T} E\left[\underline{F}_t \underline{F}_t'\right] \left(\frac{\Gamma'\Gamma}{N_1}\right)^{1/2}.$$

Although Theorem 3 shows that, without further identifying assumptions, we can only estimate the factors $\underline{F}_t$ consistently up to an invertible $Kp \times Kp$ matrix transformation, this result turns out to be sufficient for us to estimate the h-step ahead predictor consistently. More specifically, in Appendix C of the Technical Appendix, we show that, for an h-step ahead forecast, the (infeasible) forecasting equation implied by the FAVAR model (1) has the form

$$Y_{t+h} = \beta_0 + B_1'\underline{Y}_t + B_2'\underline{F}_t + \eta_{t+h}, \tag{26}$$

where $\underline{Y}_t$ and $\underline{F}_t$ are as defined in expression (4) above and where:

$$\begin{aligned} \beta_0 &= \sum_{j=0}^{h-1} J_d A^j \alpha,\ B_1' = J_d A^h \mathcal{P}'_{(d+K)p} S_d,\ B_2' = J_d A^h \mathcal{P}'_{(d+K)p} S_K \text{ and} \\ \eta_{t+h} &= \sum_{j=0}^{h-1} J_d A^j J'_{d+K} \varepsilon_{t+h-j}. \end{aligned} \tag{27}$$

Here, α and A are, respectively, the intercept (vector) and the coefficient matrix of the companion form defined in expression (5) above, $\mathcal{P}_{(d+K)p}$ is a permutation matrix such that $\mathcal{P}_{(d+K)p}\underline{W}_t = \begin{pmatrix} \underline{Y}_t \\ \underline{F}_t \end{pmatrix}$, $S_d = \begin{pmatrix} I_{dp} \\ 0_{Kp\times dp} \end{pmatrix}$, $S_K = \begin{pmatrix} 0_{dp\times Kp} \\ I_{Kp} \end{pmatrix}$, $\underset{d\times(d+K)p}{J_d} = \begin{bmatrix} I_d & 0 & \cdots & 0 \end{bmatrix}$, and $\underset{(d+K)\times(d+K)p}{J_{d+K}} = \begin{bmatrix} I_{d+K} & 0 & \cdots & 0 \end{bmatrix}$. See the beginning of Appendix C for a derivation of the equation given in expression (26). The reason expression (26) is called an infeasible forecasting equation is because $\underline{F}_t$ is not observed, so to obtain a feasible version of this forecasting equation, we must replace $\underline{F}_t$ in equation (26) with the estimate $\widehat{\underline{F}}_t$ given in expression (25). Doing so, we arrive at a feasible h-step ahead forecasting equation of the form:

$$\begin{aligned} Y_{t+h} &= \beta_0 + \sum_{g=1}^{p} B_{1,g}' Y_{t-g+1} + \sum_{g=1}^{p} B_{2,g}' \widehat{F}_{t-g+1} + \widehat{\eta}_{t+h} \\ &= \beta_0 + B_1' \underline{Y}_t + B_2' \widehat{\underline{F}}_t + \widehat{\eta}_{t+h}, \end{aligned} \tag{28}$$

where $\widehat{\eta}_{t+h} = \eta_{t+h} - B_2' \left(\widehat{\underline{F}}_t - \underline{F}_t \right)$, with $\eta_{t+h} = \sum_{j=0}^{h-1} J_d A^j J'_{d+K} \varepsilon_{t+h-j}$.

One can interpret expression (28) as a "reduced form" formulation of the forecasting equation where the reduced form parameters β_0, B_1, and B_2 are nonlinear functions of the parameters $(\mu, A_1, ..., A_p)$ of the FAVAR model, in the case where $h > 1$. For forecasting purposes, while it is possible to estimate the conditional mean of the forecasting equation (28) by estimating the underlying parameters directly using nonlinear least squares, here we choose instead to estimate the conditional mean by estimating the reduced form parameters β_0, B_1, and B_2 via linear least squares. An important reason why we choose this approach is due to complications that arise both because we are forecasting with a FAVAR which contains unobserved factors that must first be estimated and because we only make enough identifying assumptions so that the factors can only be estimated consistently up to an invertible $Kp \times Kp$ matrix transformation. In fact, it turns out that estimating the underlying parameters $\mu, A_1, ..., A_p$ by nonlinear least squares and constructing an estimator of the conditional mean of the forecasting equation based on these estimates will not lead to a consistently estimated h-step predictor, unless further identifying assumptions are made.

On the other hand, as we show in Theorem 4 below, estimating the reduced form parameters β_0, B_1, and B_2 by linear least squares allows us to construct a consistent estimator of the conditional mean, even in the absence of additional identifying assumptions. More precisely, let $\widehat{\underline{F}}_t$ denotes the factor estimates given in expression (25). Our procedure minimizes the least squares criterion function:

$$\begin{aligned} Q\left(\beta_0, B_1, B_2\right) &= \sum_{t=p}^{T-h} \left\| Y_{t+h} - \beta_0 - B_1' \underline{Y}_t - B_2' \widehat{\underline{F}}_t \right\|_2^2 \\ &= \sum_{t=p}^{T-h} \left\| Y_{t+h} - \beta_0 - \sum_{g=1}^{p} B_{1,g}' Y_{t-g+1} - \sum_{g=1}^{p} B_{2,g}' \widehat{F}_{t-g+1} \right\|_2^2 \end{aligned} \tag{29}$$

with respect to the parameters β_0, B_1, and B_2, and delivers the OLS estimates $\widehat{\beta}_0$, $\widehat{B}_1$, and $\widehat{B}_2$. We then forecast Y_{T+h} using the h-step predictor:

$$\widehat{Y}_{T+h} = \widehat{\beta}_0 + \widehat{B}_1' \underline{Y}_T + \widehat{B}_2' \widehat{\underline{F}}_T. \tag{30}$$

The following result shows that $\widehat{Y}_{T+h}$ is a consistent estimator of the conditional mean of the infeasible forecast equation (26).

Theorem 4: *Let $\widehat{Y}_{T+h}$ be as defined in expression (30). Suppose that Assumptions 2-1, 2-2, 2-3, 2-4, 2-5, 2-6, 2-7, 2-8, 2-9, 2-10, and 2-11* hold. Then,*

$$\widehat{Y}_{T+h} - \left(\beta_0 + B_1' \underline{Y}_T + B_2' \underline{F}_T\right) \xrightarrow{p} 0 \text{ as } N_1, N_2, T \to \infty.$$

4 Monte Carlo Experiment

In this section, we report some simulation results on the finite sample performance of our variable selection procedure. The model used in the Monte Carlo study is the following tri-variate FAVAR(1) process:

$$W_t = \mu + A W_{t-1} + \varepsilon_t, \tag{31}$$

$$Z_t = \gamma F_t + u_t, \tag{32}$$

where

$$W_t = \begin{pmatrix} Y_{1t} \\ Y_{2t} \\ F_t \end{pmatrix}, \mu = \begin{pmatrix} 2 \\ 1 \\ 2 \end{pmatrix}, A = \begin{pmatrix} 0.9 & 0.3 & 0.5 \\ 0 & 0.7 & 0.1 \\ 0 & 0.6 & 0.7 \end{pmatrix}, \text{ and } \gamma = \begin{pmatrix} \iota_{N_1} \\ 0 \\ {}_{N_2 \times 1} \end{pmatrix},$$

with ι_{N_1} denoting an $N_1 \times 1$ vector of ones. We consider different configurations of N, N_1, and T, as given below. For the error process in equation (31), we take $\{\varepsilon_t\} \equiv i.i.d.N(0, \Sigma_\varepsilon)$, where:

$$\Sigma_\varepsilon = \begin{pmatrix} 1.3 & 0.99 & 0.641 \\ 0.99 & 0.81 & 0.009 \\ 0.641 & 0.009 & 5.85 \end{pmatrix}.$$

The error process, $\{u_{it}\}$, in equation (32) is allowed to exhibit both temporal and cross-sectional dependence and also conditional heteroskedasticity. More specifically, we let $u_{it} = 0.8u_{it-1} + \zeta_{it}$, and following the approach for modeling cross-sectional dependence given in the Monte Carlo design of Stock and Watson (2002a), we specify: $\zeta_{it} = \left(1 + b^2\right)\eta_{it} + b\eta_{i+1,t} + b\eta_{i-1,t}$, and set $b = 1$. In addition, $\eta_{it} = \omega_{it}\xi_{it}$, with $\{\xi_{it}\} \equiv i.i.d.N(0,1)$ independent of $\{\varepsilon_t\}$, and ω_{it} follows a GARCH(1,1) process given by: $\omega_{it}^2 = 1 + 0.9\omega_{it-1}^2 + 0.05\eta_{it-1}^2$. To study the effects of varying the tuning parameter, we consider specifications where $\varphi = (\ln\ln N)^{-\vartheta}$ for $\vartheta = 0.1, 0.5, 1$ and also $\varphi = N^{-\vartheta}$ for $\vartheta = 0.2, 0.4, 0.6$.[10] We also attempt to shed light on the effects of using blocks of different sizes on the performance of our procedure. To do this, for $T = 100$, we set $\tau_1 = 2$, 3, 4, and 5; for $T = 200$, we set $\tau_1 = 5$, 6, 8, and 10; and for $T = 600$, we set $\tau_1 = 6$, 8, 10, and 12. Due to space considerations, we only report Monte Carlo results for the statistic $\sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$. Simulation results for the statistic $\max_{1\leq\ell\leq d}\left|S_{i,\ell,T}\right|$ have also been obtained by the authors. The results for $\max_{1\leq\ell\leq d}\left|S_{i,\ell,T}\right|$ are qualitatively similar to those given here for $\sum_{\ell=1}^{d} \varpi_\ell \left|S_{i,\ell,T}\right|$, and they are available from the authors upon request. In addition, since $d = 2$ in our Monte Carlo setup, we set $\varpi_1 = \varpi_2 = 1/2$. Results are gathered in Table 1 in the back of the paper. There, the acronym FPR denotes the "False Positive Rate" or the "Type I" error rate, i.e., the proportion of cases where an irrelevant variable Z_{it}, with associated coefficient $\gamma_i = 0$ is erroneously selected as a relevant variable. FNR denotes the "False Negative Rate" or the "Type II" error rate, i.e., the proportion of cases where a relevant variable is erroneously identified as being irrelevant.

Looking across each row of the table, note that FPRs decrease when moving from left to right, whereas FNRs increase. This is not surprising, because moving from $\varphi = (\ln\ln N)^{-0.1}$ to

[10] We have also obtained simulation results for the cases where $\varphi = (\ln N)^{-\vartheta}$ for $\vartheta = 0.1, 0.5, 1$ and where $\varphi = N^{-\vartheta}$ for $\vartheta = 0.3, 0.5$. The results obtained for these cases are qualitatively similar to the results reported in this paper. Hence, due to space considerations, we do not report these results here, but they are available from the authors upon request.

$\varphi = N^{-0.6}$ for a given N results in smaller values of the tuning parameter φ, and the specified threshold $\Phi^{-1}\left(1 - \frac{\varphi}{2N}\right)$ thus becomes larger. Overall, these results indicate that choosing φ in the range between $(\ln \ln N)^{-0.1}$ and $N^{-0.4}$ leads to very good performance, since within this range, neither FPR nor FNR exceeds 0.1 in any of the cases studied here. In fact, both are smaller than 0.05 in a vast majority of the cases. In contrast, choosing $\varphi = N^{-0.6}$ can lead to high FNRs, as such a choice of φ can set our threshold at such a high level that our procedure ends up having very little power.

Looking down the columns of the table, note that FPR tends to increase as τ_1 increases, whereas FNR tends to decrease as τ_1 increases. As an explanation for this result, note first that the smaller is τ_1 relative to τ, the larger is τ_2 (since $\tau = \tau_1 + \tau_2$), and thus the larger is the number of observations removed when constructing the self-normalized block sums. Intuitively, this can lead to better accommodation of the effects of dependence and better moderate deviation approximations under the null hypothesis, resulting in a lower FPR. However, removal of a larger number of observations can also lead to a reduction in power, when the alternative hypothesis is correct, so that a negative consequence of having a smaller τ_1 relative to τ is that FNR will tend to be higher in this case. The opposite, of course, occurs when we try to specify a larger τ_1 relative to τ.

Our results also show that when the sample sizes are large enough such as the cases presented in the last panel of the table, where $T = 600$ and $N = 1000$, then both FPR and FNR are very close to zero for all of the cases that we consider. Moreover, even in the case where $T = 100$ and $N = 100$, FPR and NPR rates are usually less than 0.05, and are often much smaller than that. This is in accord with the results of our theoretical analysis, which shows that our variable selection procedure is completely consistent in the sense that both the probability of a false positive and the probability of a false negative approach zero, as the sample sizes go to infinity.

5 Forecasting Experiments

In this section, we carry out prediction experiments using two different datasets and using the variable selection methodology discussed above. Our goal is to compare different variable selection methods, and more broadly to assess whether reducing the number of variables prior to factor estimation using the consistent selection and estimation technique discussed above can result in improved forecast performance.

5.1 Empirical Illustration 1 - FRED-MD Dataset

In this illustration, we forecast eight target variables from the monthly real-time macroeconomic FRED-MD dataset maintained by the St. Louis Federal Reserve Bank.[11] We follow the data cleaning methods outlined on the FRED-MD data website, as well as removing all discontinued series yielding a dataset, $\mathbf{X}$, containing 96 variables for the period 1975:1 to 2024:6. The full list of all macroeconomic variables and their transformations is available upon request from the authors.[12]

Of note is that the FRED-MD dataset used in this illustration is "truly" real-time. Consider the value of industrial production for January 2020. In February 2020, the government reported a "first release" value for January. In March 2020, they updated their "estimate" of industrial production for January. Namely, they reported a "second release" for January. This process of revision continues indefinitely. Namely, as the government changes data collection and processing methodology, collects new data and/or revises definitions of variables, new releases are reported. A "vintage" of data consists of all of the historical data that were available, in real-time, at a particular calendar date, say February 2020. This means that there is a unique vintage of industrial production data available each month, and the values of the calendar dated observations in each vintage may change over time. Using this type of data allows the practitioner to truly simulate a forecasting environment in which models are updated at each point in time using data that were actually available at that time. For further discussion of the structure of real-time datasets, as well as methods for real-time forecasting, refer to Swanson (1996), Swanson and van Dijk (2006), and Kim and Swanson (2018).

Our forecasting experiment is carried out as follows. All forecasts that we construct utilize the most recent vintage of data available. Thus, at each point in time prior to the construction of each new forecast, a new vintage of data is used for variable selection, factor construction, and forecast model estimation. The eight target variables for which we construct predictions are summarized in Table 2, and include Industrial Production (INDPRO), Civilian Unemployment Rate (UNRATE), Housing Starts: new, privately owned (HOUST), Housing Permits: new, privately owned (PERMIT), Real M2 Money Stock (M2REAL), 10-Year Government Treasury Bond Rate (GS10), CPI - All Items (CPI), S&P Common Stock Price Index - Composite (S&P500).

[11] see https://www.stlouisfed.org/research/economists/mccracken/fred-databases

[12] We use the variable transformations recommended on the FRED-MD website (see the FRED MD appendix on that website for complete details), with three exceptions. First, we do not difference unemployment. Second, we do not difference interest rate variables. Third, we do not take second differences of any variables. Instead, we take first differences of the small number of variables that are twice differenced in the list of recommended variable transformations referred to above. Tabulated results based on the original FRED-MD variable transformations are available upon request from the authors. Also, note that the original FRED-MD dataset actually begins in 1973:3.

We estimate the following forecasting model:

$$y_{t+h} = \alpha + \beta_h(L)y_t + \gamma_h(L)F_{\boldsymbol{t}} + \epsilon_{t+h}, \tag{33}$$

where y_t is the scalar target variable to be predicted, $\beta_h(L)$ and $\gamma_h(L)$ are finite order lag polynomials, $F_{\boldsymbol{t}}$ is a vector of estimated factors, and ϵ_t is a stochastic disturbance term. Lags in this model are selected using the Schwarz Information Criterion (SIC), and our benchmark model sets the lag polynomial $\gamma_h(L) = 0$. In implementing our forecast experiments, we take the initial training sample to be the period 1975:1-1999:12; but as we move forward in time with each new prediction, we use a new release of the data (i.e., an updated data vintage) in constructing our forecast. The tuning parameters as well as the number of factors are all re-estimated based on the updated data vintage. Applying our procedure to the FRED-MD dataset in this way yields estimates of N_1 ranging from 41 to 72 across the different training samples that result from our updating of the data vintage and across our 8 target variables. Since there are $N = 95$ total number of possible predictor variables excluding the target variable of interest, our estimates of N_1 imply that the number of irrelevant variables N_2 is estimated to be in the range from 24 to 54 for the range of cases described above, so that, for the FRED-MD dataset, the number of irrelevant variables N_2 seems to be a significant proportion of the total number of available variables. Given the non-triviality of N_2, having a variable selection procedure to identify those irrelevant variables which only contribute noise to the factor estimation process can potentially be beneficial both for estimation and for forecasting purposes. Moreover, as we mentioned previously in section 2, to obtain a consistent estimator $\widehat{N}_1$ of N_1, we need a completely consistent variable selection procedure, except in the case where $N_2/N_1 \to 0$. However, in light of what we have just discussed, the rate condition $N_2/N_1 \to 0$ does not seem to be reasonable for the FRED-MD dataset. Hence, if an applied researcher wishes to use an estimate of N_1 to assess the overall pervasiveness of the factors for a particular dataset of interest; then, it would make sense for her/him to use a completely consistent variable selection method since a selection method which is not completely consistent is not guaranteed to provide a consistent estimator of N_1. What still remains to be seen, however, is whether our completely consistent variable selection method will lead to improved forecast performance.

In the sequel, we carry out variable selection and dimension reduction as follows in order to estimate $F_{\boldsymbol{t}}$:[13]

Principal Components Analysis (PCA): Utilize $\mathbf{X}$ to estimate latent factors, $F_{\boldsymbol{t}}$ using PCA, with the number of factors determined using the PC_{p2} criterion in Bai and Ng (2002). The maximum number of the factors is set equal to both 4 and 8 (resulting in two distinct sets of empirical results),

[13] Note that all forecasting models are estimated using least squares, once the factors are first estimated.

following the findings of McCracken and Ng (2016), who introduce and examine the dataset that we utilize in our analysis.

Hard Thresholding (HT): For each variable in $\mathbf{X}$, and forecast horizon, h, perform a regression of y_{t+h} on lags of y_t and on $X_{i,t}$, where $X_{i,t}$ is a scalar variable in $\mathbf{X}$, for $i = 1, ..., N$, and lags of y_t are selected using the SIC. Let t_i denote the t statistic associated with X_{it-h} in the regression, and select variables, X_{it} if $|t_i| > 1.28$. If the number of selected variables is greater than 20, utilize PCA to estimate factors for inclusion in the above forecasting equation, otherwise use the AR(SIC) model. As models are re-estimated at each point in time, this approach is a hybrid, in the sense that some models may include factors as regressors, while others may be simple AR(SIC) models. Note that for all variables except the S&P500, the thresholding model was replaced with the AR(SIC) benchmark for less than 10% of the total number of forecasting periods.

Chao-Swanson Variable Selection (CS): Use the self-normalized statistic given in expressions (21) and (22) above for variable selection, and then estimate factors for inclusion in the forecasting equation using PCA. Consider the following sets of tuning parameter values: $\{\tau = 5, \tau_1 = 3, 5\}$ and $\{\tau = 10, \tau_1 = 6, 8\}$, with

$$\varphi = \begin{cases} (lnlnN)^{-0.1} & (lnlnN)^{-0.6} & (lnN)^{-0.1} & (lnN)^{-0.6} & N^{-0.1} & N^{-0.6} \\ (lnlnN)^{-0.2} & (lnlnN)^{-0.7} & (lnN)^{-0.2} & (lnN)^{-0.7} & N^{-0.2} & N^{-0.7} \\ (lnlnN)^{-0.3} & (lnlnN)^{-0.8} & (lnN)^{-0.3} & (lnN)^{-0.8} & N^{-0.3} & N^{-0.8} \\ (lnlnN)^{-0.4} & (lnlnN)^{-0.9} & (lnN)^{-0.4} & (lnN)^{-0.9} & N^{-0.4} & N^{-0.9} \\ (lnlnN)^{-0.5} & (lnlnN)^{-1} & (lnN)^{-0.5} & (lnN)^{-1} & N^{-0.5} & N^{-1} \end{cases}.$$

Tuning parameters used for each value of h and target variable are selected in real-time prior to the construction of each new forecast by using an initial "training dataset" consisting of the first 25 years of data. This training dataset is partitioned into an in-sample period of 20 years and an out-of-sample period of 5 years. In subsequent prediction experiments, tuning parameter is set equal to that yielding the smallest mean square forecast error (MSFE) after constructing real-time predictions over this 5 year period.

In summary, we construct real-time h-month ahead predictions using monthly data, with h =1, 3, 6, and 12. Our forecasting models are called PCA, CS, and HT, in reference to the manner in which variable selection prior to factor estimation is carried out. The sample period used in our analysis is 1975:1-2024:6, and our ex ante forecast period is 2000:1-2024:6. Our initial training sample period is 25 years from 1975:1 to 1999:12, as discussed above, and both rolling and recursive windows of data are used when re-estimating models at each point in time. Forecasting performance is evaluated using point mean squared forecast errors (MSFEs), where MSFE=$\frac{1}{P}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2$,

and $\hat{y}_t$ denotes the real-time prediction for target variable y_t. In our tabulated results, MSFEs, relative to those of the benchmark AR(SIC) model are reported. Additionally, we report the results of Giacomini-White (GW) tests (see Giacomini and White (2006)), which can be viewed as conditional Diebold-Mariano (DM) predictive accuracy tests (see Diebold and Mariano (1995)). Recall that the null hypothesis of the DM test when formulated using the conditioning approach of Giacomini and White is: $H_0 : \mathrm{E}[L(\hat{\epsilon}^{(1)}_{t+h})|G_t] - \mathrm{E}[L(\hat{\epsilon}^{(2)}_{t+h})|G_t] = 0$, where the $\hat{\epsilon}^{(i)}_{t+h}$ are prediction errors associated with model i, for $i = 1, 2$, and G_t denotes the conditioning set, which includes the model and estimated parameters. Here, $L(\cdot)$ is a quadratic loss function, and the test statistic is $\mathrm{DM}_P = P^{-1} \sum_{t=1}^{P} \frac{d_{t+h}}{\hat{\sigma}_{\bar{d}}}$, where $d_{t+h} = [\hat{\epsilon}^{(1)}_{t+h}]^2 - [\hat{\epsilon}^{(2)}_{t+h}]^2$, $\bar{d}$ denotes the mean of d_{t+h}, $\hat{\sigma}_{\bar{d}}$ is a heteroskedasticity and autocorrelation consistent estimate of the standard deviation of $\bar{d}$, and P denotes the number of ex-ante predictions used to construct the test statistic.[14] If the null hypothesis is rejected and the relative MSFE is greater than 1, then the AR benchmark is preferred.

Our main empirical findings are gathered in Table 3 and Table 4. In these tables, all entries are relative MSFEs, with our AR(SIC) benchmark in the denominator. Additionally, bolded entries indicate the "MSFE-best" method for a particular target variable, forecast horizon, and estimation window type (i.e. rolling or recursive). Starred entries denote rejection of the null hypothesis of equal forecast accuracy when comparing the listed model against the AR(SIC) benchmark. Note that Table 3 reports the relative MSFE results for the case where in choosing the number of factors using the PC_{p2} criterion in Bai and Ng (2002), we set the maximum allowable number of factors to be 4, whereas Table 4 reports the results for the case where the maximum allowable number of factors is 8. Moreover, the first three columns of results given in Tables 3 and 4 are for the case where the estimation is conducted using a recursive data window, while the last three columns of results in both tables are for the case where estimation is conducted using a rolling data window. A number of conclusions can be drawn from examining the results in these tables. Focusing on Table 3, we see that when a recursive data window is used, then we see that CS has a smaller MSFE than both PCA and HT in 16 out of the 32 possible cases, when comparing the MSFE results of PCA, HT, and CS across 8 different target variables and 4 different forecast horizons. On the other hand, HT is the top performer in 11 out of 32 cases while PCA wins in 5 out of 32 cases. The results are even better when we adopt a rolling data window. In this case, CS beats both HT and PCA in 24 out of 32 cases, whereas HT wins in 5 out of 32 cases and PCA only wins in 3 out of 32 cases. Turning our attention to Table 4, we see that in the case where a recursive data window is used, CS beats both HT and PCA in 18 out of 32 cases, while HT wins in 7 out of 32 cases and PCA wins in 6 out of 32 cases, with HT and PCA sharing first place in the one remaining case. Moreover,

[14] In this paper, we report test results for the Wald version of this test statistic (see Giacomini and White (2006) for further details).

the results reported in Table 4 for the case where a rolling data window is used find CS winning in 19 out of 32 cases, HT winning in 7 out of 32 cases, and PCA winning in 6 out of 32 cases. Finally, if we compare CS directly with PCA across 8 different target variables, 4 forecast horizons, 2 choices of data window specification (recursive or rolling), and 2 choices of maximum allowable number of factors (4 or 8); we see that CS outperforms PCA in 97 out of the 128 possible cases considered here, whereas PCA only outperforms CS in 31 cases. The latter results also provide some additional evidence that the inclusion of the irrelevant variables actually has a non-negligible influence on our estimation and forecasting results; since, if the number of irrelevant variables N_2 is negligibly small so that their inclusion has only a trivial impact; then, we would expect more balance with respect to the number of cases for which PCA beats CS versus the number of cases for which CS beats PCA.

To summarize, overall CS wins more frequently than either HT or PCA. However, note that no method uniformly beats all of the competing methods across all settings, which should not be surprising when one applies econometric methods to real world data. The goal of this empirical illustration is simply to provide evidence that the CS variable selection procedure introduced in this paper can be useful for FRED-MD forecasting. Needless to say, our results are not meant to imply that the other methods are not also useful. Rather, we believe that the CS variable selection method should be added to the "toolbox" of methods used when constructing macroeconomic forecasts.

5.2 Empirical Illustration 2 - Global VAR Modelling Dataset

In this illustration, we forecast 5 target variables from the quarterly real-time macroeconomic Global VAR Modelling (GVAR) dataset maintained by L. Vanessa Smith.[15] The GVAR dataset extends the original one used in Dees, di Mauro, Pesaran and Smith (2007), which covers the period 1979:Q1-2003:Q4 (see also Pesaran, Schuermann and Smith (2009)). More specifically, the dataset includes 4 vintages of real-time data on 6 variables for 33 countries. The four vintages include data collected in 2011 for the period 1979:Q1-2011:Q2, data collected in 2013 for the period 1979:Q1-2013:Q1, data collected in 2017 for the period 1979:Q1-2016:Q4, and data collected in 2023 for the period 1979:Q1-2023:Q3. The 5 variables that we forecast include GDP growth, inflation, equity returns, short-term interest rates, and long-term interest. These variables are predicted for 6 different countries, including the USA, the UK, Germany, France, Italy, and Japan. Additionally, the GVAR dataset to which we apply the CS and HT variable selection methods prior to constructing factors includes 6 variables for each of 33 countries for a total of 198 variables. The conventional PCA procedure does not involve variable pre-screening, so it uses this entire set of

[15] see https://sites.google.com/site/gvarmodelling/home

variables (again excluding the target variable of interest) in estimating the factors.

When constructing predictions, we use a variant of the forecasting model analyzed in the previous empirical illustration, specified as follows:

$$y_{t+h} = \alpha + \beta_h(L)y_t + \gamma_h(L)F_{\boldsymbol{t}} + \delta_h(L)W_{\boldsymbol{t}} + \epsilon_{t+h}. \tag{34}$$

All terms in the above equation are defined in the previous empirical illustration, with the exception of $\delta_h(L)W_{\boldsymbol{t}}$, where $W_{\boldsymbol{t}}$ contains the 5 target macroeconomic variables for each of the 6 countries in our analysis, and $\delta_h(L)$ is a conformably defined lag polynomial. The forecasting component of this illustration follows the approach of Pesaran, Schuermann and Smith (2009), where forecasts for the last 8 periods of our sample period are constructed, and models are ranked based on comparison of MSFEs that are constructed by averaging ranks across all 5 variables. Our sample period is 1979:Q1-2023:Q3, so that only the latest vintage of data is used in our analysis, and the ex ante forecasting period is 2021:Q4-2023:Q3.[16] The models that we use when constructing include PCA, CS, HT, PCA+macro, CS+macro, and HT+macro. The first three models (i.e., PCA, CS, and HT) are defined above. The latter three models include lags of the 5 macroeconomic variables as explanatory variables, rather than just lags of the target variable, with lags selected using the SIC. All models are estimated recursively, prior to the construction of each new forecast, using the same procedures described in the previous section of this paper. Additionally, models are estimated using PCA, followed by least squares, with lags selected using the SIC and the number of factors selected using the PC_{p2} criterion of Bai and Ng (2002).

A summary of our results is presented in Table 5, where the average model ranks are reported. These ranks are determined by comparing the MSFEs produced by the different methods for a given country, target variable, and forecast horizon. Model ranks are defined as "1" for the lowest MSFE model, "2" for the second "lowest" MSFE model, etc. For example, if a model, say PCA, yields the lowest MSFE for French GDP growth at the h=1 step ahead horizon, PCA is assigned a "1" as it is the lowest MSFE model for that particular country, horizon, and variable permutation.

[16] Note that Pesaran, Schuermann and Smith (2009) use an earlier vintage of data in their forecasting analysis, and not the more recently available vintage ending in 2023:Q3 that we use. Additionally, their analysis and ours do not use a new vintage of data for the construction of each new forecast, and hence are not truly real-time in the sense discussed in our previous empirical illustration. For this reason, we also carried out a alternative "real-time" version of the experiment reported on in this section by constructing 1 to 4 step ahead forecasts using each of the first three vintages of data in the GVAR dataset (the 4th vintage of data was used as the "fully revised" data against which all forecasts were compared when constructing forecast errors). More specifically, the 1979:Q1-2011:Q2 vintage of data was used to construct h=1,...,4 step ahead forecasts beginning with the period 2011:Q3, the 1979:Q1-2013:Q1 vintage of data was used to construct h=1,...,4 step ahead forecasts beginning with the period 2013:Q2, and the 1979:Q1-2016:Q4 vintage of data was used to construct h=1,...,4 step ahead forecasts beginning with the period 2017:Q1. Results from this experiment yield rankings that are identical to those reported on in the sequel, and are available upon request from the authors.

Just as is done in Pesaran, Schuermann and Smith (2009), the model rankings reported in Table 5 average the individual rankings across all 5 forecasting variables. For example, note that the first numerical entry in the Table 5 is 2.4. This means that for the USA, on average, the PCA model ranks 2.4, when averaging the rank of the PCA model across all five of the variables that we forecast.

A number of conclusions can be made by examining the results provided in Table 5. First, in the top panel of Table 5, which summarizes results for $h = 1$, we see that the CS model is the top performer for 2 of 6 countries (i.e., Germany and Japan). This increases to 4 of 6 countries when counting the number of times that the CS model "wins" or comes in second place, with the USA and France constituting the additional two countries. For the 2-step ahead horizon (refer to Panel 2), the CS model "wins" for 4 of 6 countries, and the CS+macro model wins for 1 country. For this forecast horizon, the only country for which the top performer is not a CS-type model is Germany, as HT outperforms all other methods in forecasting German data in the $h = 2$ case. The results are **roughly** similar when assessing the number of CS model wins for the additional two forecast horizons, so that overall the CS model is the top performer in terms of the number of times it achieves the highest average ranking. Second, the model which comes in second place in terms of the number of times that it achieves the highest ranking is PCA. More specifically, PCA "wins" in 8 of 24 cases and also ties for first place with HT in one other case, when comparing results across all 6 countries and 4 forecast horizons. For comparison, note that the two CS-type models (i.e., CS and CS+macro) together "win" in 12 of 24 cases, with the CS model winning in 11 of 24 cases and the CS+macro model winning in one additional case. Moreover, if we were to compare CS directly with PCA across results given for the 6 different countries and 4 different forecast horizons, we see that CS outperforms PCA in terms of average rank in 14 out of the 24 possible cases. Third, observe that our models which include additional macroeconomic explanatory variables do not yield superior predictions, when compared against more parsimonious models that include only factors and lags of the target variable being predicted. This is as expected, given how heavily parameterized our models with additional explanatory variables are. In summary, this experiment yields further evidence that the CS variable selection method is useful when specifying commonly used factor augmented (vector) autoregression type time series models.

6 Conclusion

In this paper, we present a novel and completely consistent approach for variable selection in high dimensional factor estimation problems. Our method can be useful in contexts where not all available variables actually load on the underlying factors so that the variables which do not load

can be considered to be irrelevant in the sense that they contribute only noise and not signal to the factor estimation process. We show that our variable selection procedure allows for the consistent estimation of the conditional mean of a factor-augmented forecasting equation based on a FAVAR model, even in cases where the number of irrelevant variables may be quite substantial. Our new variable selection procedure is based on a self-normalized score statistic, and it correctly identifies the set of variables which load significantly on the underlying factors, with probability approaching one, as the sample sizes go to infinity. Our theoretical analysis suggests that our method for variable selection may be a useful complement to extant methods for pre-screening variables prior to latent factor estimation. Some support for this conclusion is also given in the form of a Monte Carlo experiment indicating that the variables selection method has very small false positive and false negative rates even for samples that are not very large. In addition, we present two empirical illustrations which show the good forecast performance of our methodology when compared to the conventional PCA and hard thresholding procedures for variable selection.

References

[1] Andrews, D.W.K. (1984): "Non-strong Mixing Autoregressive Processes," *Journal of Applied Probability*, 21, 930-934.

[2] Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

[3] Bai, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica*, 71, 135-171.

[4] Bai, J. and S. Ng (2002): "Determining the Number of Factors in Approximate Factor Models," *Econometrica*, 70, 191-221.

[5] Bai, J. and S. Ng (2008): "Forecasting Economic Time Series Using Targeted Predictors," *Journal of Econometrics*, 146, 304-317.

[6] Bai, J. and S. Ng (2023): "Approximate Factor Models with Weaker Loading," *Journal of Econometrics*, 235. 1893-1916.

[7] Bai, Z. D. and Y. Q. Yin (1993): "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix," *Annals of Probability*, 21, 1275-1294.

[8] Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2006): "Prediction by Supervised Principal Components," *Journal of the American Statistical Association*, 101, 119-137.

[9] Barshan, E., A. Ghodsi, Z. Azimifar, and M.Z. Jahromi (2011): "Supervised Principal Component Analysis: Visualization, Classification and Regression on Subspaces and Submanifolds," *Pattern Recognition*, 44, 1357-1371.

[10] Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369-2429.

[11] Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *Review of Economic Studies*, 81, 608-650.

[12] Belloni, A., V. Chernozhukov, and L. Wang (2014): "Pivotal Estimation via Square-Root Lasso in Nonparametric Regression," *Annals of Statistics*, 42, 757-788.

[13] Billingsley, P. (1995): *Probability and Measure*. New York: John Wiley & Sons.

[14] Borovkova, S., R. Burton, and H. Dehling (2001): "Limit Theorems for Functionals of Mixing Processes to U-Statistics and Dimension Estimation," *Transactions of the American Mathematical Society*, 353, 4261-4318.

[15] Breiman, L. (1995). "Better Subset Regression Using the Nonnegative Garrote," *Technometrics*, 37, 373-384.

[16] Breiman, L. (1996). "Bagging Predictors," *Machine Learning*, 26, 123-140.

[17] Breiman, L. (2001). "Random Forests," *Machine Learning*, 45, 5-32.

[18] Chao, J. C., K. Qiu, and N. R. Swanson (2023): "Consistent Factor Estimation and Forecasting in Factor-Augmented VAR Models," University of Maryland and Rutgers University Working Paper.

[19] Carrasco, M. and B. Rossi (2016): "In-sample Inference and Forecasting in Misspecified Factor Models," *Journal of Business & Economic Statistics,* 34, 313-338.

[20] Chen, X., Q. Shao, W. B. Wu, and L. Xu (2016): "Self-normalized Cramér-type Moderate Deviations under Dependence," *Annals of Statistics*, 44, 1593-1617.

[21] Davidson. J. (1994): *Stochastic Limit Theory: An Introduction for Econometricians*. New York: Oxford University Press.

[22] Davidson, K. R. and S. J. Szarek (2001): "Local Operator Theory, Random Matrices and Banach Spaces." In *Handbook of the Geometry of Banach Spaces*, 1, 317-366. Amsterdam: North-Holland.

[23] Dees, S., F. Di Mauro, M.H. Pesaran, and L.V. Smith (2007): "Exploring the International Linkages of the Euro Area: A Global VAR Analysis," *Journal of Applied Econometrics,* 22, 1-38.

[24] Diebold, F.X. and R.S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 20, 134-144.

[25] Fan, J., Y. Ke, and Y. Liao (2021): "Augmented Factor Models with Applications to Validating Market Risk Factors and Forecasting Bond Risk Premia," *Journal of Econometrics*, 222, 269-294.

[26] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005): "The Generalized Dynamic Factor Model, One-Sided Estimation and Forecasting," *Journal of the American Statistical Association*, 100, 830-840.

[27] Freund, Y. and R. Schapire (1997). "A Decision-Theoretic Generalization of Online Learning and an Application to Boosting, " *Journal of Computer and System Sciences*, 55, 119-139.

[28] Giacomini, R. and H. White (2006): "Tests of Conditional Predictive Ability," *Econometrica*, 74, 1545-1578.

[29] Giglio, S., D. Xiu, and D. Zhang (2023a): "Test Assets and Weak Factors," *Joural of Finance*, forthcoming.

[30] Giglio, S., D. Xiu, and D. Zhang (2023b): "Prediction When Factors Are Weak," Working Paper, Yale School of Management and the Booth School of Business, University of Chicago.

[31] Golub, G. H. and C. F. van Loan (1996): *Matrix Computations*, 3rd Edition. Baltimore: The Johns Hopkins University Press.

[32] Goroketskii, V. V. (1977): "On the Strong Mixing Property for Linear Sequences," *Theory of Probability and Applications*, 22, 411-413.

[33] Horn, R. and C. Johnson (1985): *Matrix Analysis*. Cambridge University Press.

[34] Johnstone, I. M. and A. Lu (2009): "On Consistency and Sparsity for Principal Components Analysis in High Dimensions," *Journal of the American Statistical Association*, 104, 682-697.

[35] Johnstone, I. M. and D. Paul (2018): "PCA in High Dimensions: An Orientation," *Proceedings of the IEEE*, 106, 1277-1292.

[36] Kelly, B. and S. Pruitt (2015): "The Three-Pass Regression Filter: A New Approach to Forecasting Using Many Predictors," J*ournal of Econometrics,* 186, 294–316.

[37] Kim, H.-H. and N.R. Swanson (2014): "Forecasting Financial And Macroeconomic Variables Using Data Reduction Methods: New Empirical Evidence," *Journal of Econometrics,* 178, 352–367.

[38] Kim, H.-H. and N.R. Swanson (2018): "Methods for Backcasting, Nowcasting and Forecasting Using Factor-MIDAS: With an Application to Korean GDP," *Journal of Forecasting*, 37, 281-302.

[39] Lee, T.-H., A. Ullah, and R. Wang 2020): "Bootstrap Aggregating and Random Forest," in *Macroeconomic Forecasting in the Era of Big Data Theory and Practice*, eds. Fuleky, P., New York: Springer.

[40] Lee, T.-H. and Y. Yang, "Bagging Binary and Quantile Predictors for Time Series", *Journal of Econometrics* 135, 465-497.

[41] Lütkepohl, H. (2005): *New Introduction to Multiple Time Series Analysis*. New York: Springer.

[42] McCracken, M.W. and S. Ng (2016): "FRED-MD: A Monthly Database for Macroeconomic Research," *Journal of Business and Economic Statistics*, 34, 574-589.

[43] Nadler, B. (2008): "Finite Sample Approximation Results for Principal Component Analysis: A Matrix Perturbation Approach," *Annals of Statistics*, 36, 2791-2817.

[44] Paul, D. (2007): "Asymptotics of Sample Eigenstructure for a Large Dimensional Spiked Covariance Model," *Statistica Sinica*, 17, 1617-1642.

[45] Pesaran, M.H., T. Schuermann, and L.V. Smith (2009): "Forecasting Economic and Financial Variables With Global VARs," *International Journal of Forecasting*, 25, 642-675.

[46] Pham, T. D. and L. T. Tran (1985): "Some Mixing Properties of Time Series Models," *Stochastic Processes and Their Applications*, 19, 297-303.

[47] Qiu, A. and Z. Qu (2021): "Modeling Regime Switching in High-Dimensional Data with Applications to U.S. Business Cycles," Working Paper, Boston University.

[48] Ruhe, A. (1975): "On the Closeness of Eigenvalues and Singular Values for Almost Normal Matrices," *Linear Algebra and Its Applications*, 11, 87-94.

[49] Shen, D., H. Shen, H. Zhu, J.S. Marron (2016): "The Statistics and Mathematics of High Dimension Low Sample Size Asymptotics," *Statistica Sinica*, 26, 1747-1770.

[50] Stewart, G.W. (1973): "Error and Perturbation Bounds for Subspaces Associated with Certain Eigenvalue Problems," *SIAM Review*, 15, 727-764.

[51] Stewart, G.W. and J. Sun (1990): *Matrix Perturbation Theory*. Boston: Academic Press.

[52] Stock, J. H. and M. W. Watson (2002a): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167-1179.

[53] Stock, J. H. and M. W. Watson (2002b): "Macroeconomic Forecasting Using Diffusion Indexes," *Journal of Business and Economic Statistics*, 20, 147-162.

[54] Swanson, N.R. (1996): "Forecasting Using First-Available Versus Fully Revised Economic Time-Series Data," *Studies in Nonlinear Dynamics and Econometrics*, 1, 47-64.

[55] Swanson, N.R. and D. van Dijk (2006): "Are Statistical Reporting Agencies Getting It Right? Data Rationality and Business Cycle Asymmetry," *Journal of Business and Economic Statistics*, 24, 24-42.

[56] Tibshirani, R. (1996): "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society, Series B*, 58, 267-288.

[57] Tu, Y. and Lee, T.-H. (2019): "Forecasting Using Supervised Factor Models," *Journal of Management Science and Engineering,* 4, 12-27.

[58] Vershynin, R. (2012): "Introduction to the Non-asymptotic Analysis of Random Matrices," In *Compressed Sensing, Theory and Applications,* 210-268. Cambridge University Press.

[59] Wang, W. and J. Fan (2017): "Asymptotics of Empirical Eigenstructure for High Dimensional Spiked Covariance," *Annals of Statistics*, 45, 1342-1374.

[60] Zou, H.D., T. Hastie, and R. Tibshirani (2006): "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics,* 15, 265-286.

[61] Zhou, Z. and X. Shao (2013): "Inference for Linear Models with Dependent Errors," *Journal of the Royal Statistical Society Series B*, 75, 323-343.

Table 1: Monte Carlo Results for Variable Selection Using $\mathbb{S}_{i,T}^{+} = \sum_{\ell=1}^{d} \varpi_\ell \, |S_{i,\ell,T}|$ Statistic*

		$N = 100$	$N_1 = 50$	$T = 100$	$\tau = 5$		
		$\varphi = (\ln \ln N)^{-0.1}$	$\varphi = (\ln \ln N)^{-0.5}$	$\varphi = (\ln \ln N)^{-1}$	$\varphi = N^{-0.2}$	$\varphi = N^{-0.4}$	$\varphi = N^{-0.6}$
$\tau_1 = 2$	FPR	0.03916	0.03350	0.02678	0.01460	0.00382	0.00076
	FNR	0.00046	0.00068	0.00104	0.00284	0.01674	0.09412
$\tau_1 = 3$	FPR	0.04544	0.03902	0.03110	0.01810	0.00526	0.00092
	FNR	0.00022	0.00032	0.00052	0.00172	0.01100	0.06942
$\tau_1 = 4$	FPR	0.05408	0.04650	0.03756	0.02224	0.00702	0.00162
	FNR	0.00016	0.00024	0.00034	0.00118	0.00828	0.05194
$\tau_1 = 5$	FPR	0.06332	0.05462	0.04558	0.02796	0.00924	0.00232
	FNR	0.00014	0.00018	0.00034	0.00084	0.00574	0.03948
		$N = 200$	$N_1 = 100$	$T = 100$	$\tau = 5$		
$\tau_1 = 2$	FPR	0.01913	0.01470	0.01068	0.00486	0.00064	0.00002
	FNR	0.00206	0.00282	0.00449	0.01415	0.09966	0.48356
$\tau_1 = 3$	FPR	0.02341	0.01842	0.01365	0.00657	0.00098	0.00005
	FNR	0.00143	0.00190	0.00315	0.00921	0.07372	0.40894
$\tau_1 = 4$	FPR	0.02869	0.02306	0.01733	0.00841	0.00133	0.00004
	FNR	0.00111	0.00145	0.00224	0.00661	0.05564	0.34279
$\tau_1 = 5$	FPR	0.03506	0.02903	0.02194	0.01124	0.00213	0.00017
	FNR	0.00086	0.00112	0.00172	0.00477	0.04258	0.28620
		$N = 400$	$N_1 = 200$	$T = 200$	$\tau = 10$		
$\tau_1 = 5$	FPR	0.00214	0.00148	0.00090	0.00030	2.5×10^{-5}	0.00000
	FNR	7.5×10^{-5}	0.00016	0.00040	0.00231	0.06894	0.67266
$\tau_1 = 6$	FPR	0.00249	0.00166	0.00104	0.00034	0.00002	0.00000
	FNR	0.00004	0.00009	0.00025	0.00148	0.05058	0.60968
$\tau_1 = 8$	FPR	0.00337	0.00235	0.00142	0.00046	0.00004	0.00000
	FNR	0.00001	0.00002	0.00008	0.00068	0.02712	0.48133
$\tau_1 = 10$	FPR	0.00484	0.00350	0.00220	0.00079	7.5×10^{-5}	5.0×10^{-6}
	FNR	0.00001	0.00001	0.00002	0.00034	0.01535	0.36382
		$N = 1000$	$N_1 = 500$	$T = 600$	$\tau = 12$		
$\tau_1 = 6$	FPR	0.00155	0.00121	0.00086	0.00038	0.00006	0.00001
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 8$	FPR	0.00201	0.00153	0.00106	0.00049	8.2×10^{-5}	1.4×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 10$	FPR	0.00274	0.00216	0.00155	0.00072	0.00016	3.2×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
$\tau_1 = 12$	FPR	0.00421	0.00332	0.00242	0.00115	0.00028	6.0×10^{-5}
	FNR	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000

* Notes: False positive and negative rates are reported for various values of N, N_1, and T. Results are based on 1000 simulations. See Section 3 for complete details.

Table 2: Empirical Illustration 1 - FRED MD Dataset: Target Forecast Variables*

Target Variable	Abbreviation	Data Transformation
Industrial Production	INDPRO	$\Delta log(y_t)$
Civilian Unemployment Rate	UNRATE	y_t
Housing Starts (new, privately owned)	HOUST	$log(y_t)$
Housing Permits (new, privately owned)	PERMIT	$log(y_t)$
Real M2 Money Stock	M2REAL	$\Delta log(y_t)$
10-Year Government Treasury Bond Rate	GS10	y_t
CPIAUCSL (all items)	CPI	$\Delta log(y_t)$
S&P Common Stock Price Index (composite)	S&P500	$\Delta log(y_t)$

* Notes: This table lists the target forecast variables that are predicted in our empirical illustration, and associated data transformations.

Table 3: Empirical Illustration 1 - FRED MD Dataset - Forecasting Results Based on the Use of Alternative Variable Selection Methods With a Maximum of 4 Factors (Forecast Period 2000:1-2024:6)*

		Estimation Uses Recursive Data Window			Estimation Uses Rolling Data Window		
	Target Variable	PCA	HT	CS	PCA	HT	CS
h=1	INDPRO	**0.924**	0.926	0.926	**0.959**	0.990	0.964
	UNRATE	0.576	0.637	**0.568**	0.423	**0.417**	0.455
	HOUST	1.002	**0.999**	1.003	0.999	**0.992**	1.043
	PERMIT	1.033	1.020	**0.973**	0.959	0.950	**0.873**
	M2REAL	**0.999**	1.020	1.001	1.094	1.021	**1.015**
	GS10	1.102	1.086	**1.065**	1.069	1.074	**1.019**
	CPI	1.003	1.004	**1.002**	1.097	**0.992**	1.031
	S&P500	1.061 **	1.111 **	**1.054 *** **	1.126	1.453 **	**1.079**
h=3	INDPRO	1.036	1.052	**1.016**	1.078	**0.964**	1.032
	UNRATE	**0.564**	0.586	0.580	0.801	0.801	**0.779**
	HOUST	0.996	**0.985**	0.988	1.004	1.005	**0.995**
	PERMIT	1.001	**0.982**	0.986	1.009	**0.995 ***	0.998
	M2REAL	1.008	1.006	**0.995**	1.075	1.060	**1.037 ***
	GS10	1.188 **	**1.135 ***	1.168	1.076	1.090	**0.962**
	CPI	0.971	0.967	**0.960**	0.976	0.970	**0.966 ***
	S&P500	1.033 ***	**1.000**	1.025 **	1.068	1.094	**1.04**
h=6	INDPRO	1.047 *	1.048	**1.031**	1.231	1.231	**1.038**
	UNRATE	0.786 ***	0.813 ***	**0.753 *** **	0.733	0.733	**0.726**
	HOUST	1.003	**0.976**	0.986	1.014	0.994	**0.980**
	PERMIT	1.014	**0.975**	0.983	1.019	0.987	**0.981**
	M2REAL	1.008	1.007	**1.000**	1.023	1.008	**1.002**
	GS10	1.175 *	**1.085**	1.144	1.092	1.099	**1.044**
	CPI	**0.995**	1.006	1.003	1.012	0.989	**0.981**
	S&P500	**1.026 ***	1.038	1.034 *	1.178	1.085	**1.042**
h=12	INDPRO	1.009	1.011	**1.001**	1.003	1.021	**1.001**
	UNRATE	0.762 ***	0.751 ***	**0.734 *** **	0.876	0.878	**0.771**
	HOUST	0.951	**0.939 ***	0.956	1.009	1.006	**0.952**
	PERMIT	0.979	0.944	**0.94**	1.017	0.954	**0.953**
	M2REAL	1.002	**0.967**	0.988	1.002	0.967	**0.96**
	GS10	1.224	1.152 *	**1.109**	**1.121**	1.151	1.187
	CPI	0.949	**0.938**	0.955	0.958	0.948	**0.931**
	S&P500	1.002	1.020	**1.000**	**0.997**	1.031	1.015

* Notes: Results reported in this table summarize findings from a prediction experiment that uses real-time data collected in the St. Louis Federal Reserve Bank's FRED-MD dataset (see https://www.stlouisfed.org/research/economists/mccracken/fred-databases) to construct real-time forecasts. Tabulated entries are relative MSFEs (where the AR(SIC) MSFE is the denominator), for forecast horizons h=1,3,6, and 12 months, and for recursive and rolling data window estiamtion schemes. Entries in bold denote lowest relative MSFEs for a given target variable, forecast horizon, and data windowing scheme. In all cases, factors are estimated using PCA, and the number of factors is estiamted using the PC_{p2} criterion of Bai and Ng (2002). The variables used in factor estimation are selected using PCA (all variables), hard thresholding (HT) and the CS test (CS). Starred entries indicate rejection of the null hypothesis of equal conditional predictive ability, at significance levels $p = 0.01$ $(***)$, $p = 0.05$, $(**)$, and $p = 0.10$ $(*)$. See Section 5.1 for complete details.

Table 4: Empirical Illustration 1 - FRED MD Dataset - Forecasting Results Based on the Use of Alternative Variable Selection Methods With a Maximum of 8 Factors (Forecast Period 2000:1-2024:6)*

		Estimation Uses Recursive Data Window			Estimation Uses Rolling Data Window		
	Target Variable	PCA	HT	CS	PCA	HT	CS
h=1	INDPRO	**0.941**	**0.941**	0.992	**0.906**	0.982	1.009
	UNRATE	**0.612**	0.617	0.704	0.518	0.528	**0.407**
	HOUST	0.998	**0.993**	1.032	1.045	**0.989**	1.037
	PERMIT	0.971	1.008	**0.941**	0.910	0.971	**0.816**
	M2REAL	1.027	**1.004**	1.008	**1.013**	1.021	1.041
	GS10	1.196	1.128	**1.115**	1.096	1.110	**1.070**
	CPI	0.968	0.980	**0.966**	1.021	**0.979**	1.035
	S&P500	1.133 ***	1.129 **	**1.094 *****	1.295	1.438	**1.151 ****
h=3	INDPRO	1.086	**1.068**	1.087	1.083	**0.955**	1.095 *
	UNRATE	**0.574**	0.579	0.575	0.811	**0.799**	0.801
	HOUST	0.969 *	0.971 *	**0.968 ***	**0.991**	1.001	0.994
	PERMIT	0.969	**0.966**	0.977	0.995	0.991	**0.976**
	M2REAL	**1.027**	1.032	1.048 **	**1.025**	1.086	1.036
	GS10	1.279	1.245 *	**1.128**	1.125	1.126	**1.077**
	CPI	0.973	0.945	**0.922 ***	**0.959**	0.984	0.990
	S&P500	1.065 **	**1.028**	1.054 **	1.095	**1.041**	1.073
h=6	INDPRO	**1.038**	1.045	1.068 **	1.187	1.189	**1.132 ***
	UNRATE	0.803 ***	0.807 ***	**0.771 *****	**0.786**	0.807 **	0.824
	HOUST	**0.965 ****	0.967 *	0.966 *	1.020	1.020	**0.983**
	PERMIT	**0.963 ***	0.966	0.975	1.005	0.996	**0.956**
	M2REAL	1.031	1.039 *	**1.001**	1.014	1.042	**1.000**
	GS10	1.305	1.33 *	**1.250**	1.281 **	1.216 *	**1.188**
	CPI	1.064	1.040	**0.988**	1.035	1.044	**1.007**
	S&P500	1.081	1.108 **	**1.043 ***	1.38 *	1.24	**1.132**
h=12	INDPRO	1.018	1.020	**1.011**	1.025	**1.012**	1.033
	UNRATE	0.74 ***	0.741 ***	**0.727 *****	0.781 *	0.745 *	**0.724**
	HOUST	0.932 **	0.921 **	**0.917 ****	1.028	0.970	**0.946 ***
	PERMIT	0.958	**0.93**	0.943	1.037	1.002	**0.931**
	M2REAL	0.987	0.972	**0.971**	1.002	**0.964**	0.977
	GS10	1.437	1.401 **	**1.308**	1.327 *	1.283	**1.277**
	CPI	0.972	0.933	**0.914 ***	0.959	0.977	**0.957**
	S&P500	1.042	**1.022**	1.024	1.018	1.060	**1.015**

* Notes: See notes to Table 3.

Table 5: Empirical Illustration 2 - Multi Country Dataset - Average Predictive Accuracy Rank Scores by Country and Forecast Horizon*

	PCA	CS	HT	PCA+macro	CS+macro	HT+macro
Average Ranks Across All Variables for Forecast Horizon h=1						
USA	2.4	3.6	3.6	3.8	3.8	3.8
UK	3.0	3.4	3.0	3.6	3.4	4.6
Germany	2.6	2.4	2.8	4.0	5.2	4.0
France	2.8	2.3	1.3	5.0	5.5	4.3
Italy	5.2	3.4	2.8	4.0	3.2	2.4
Japan	3.3	2.8	4.3	3.5	2.8	4.5
Average Ranks Across All Variables for Forecast Horizon h=2						
USA	3.0	2.4	3.4	4.2	3.6	4.4
UK	3.2	3.0	4.8	3.8	2.4	3.8
Germany	2.8	3.0	2.0	4.2	4.6	4.4
France	3.3	2.0	2.8	5.3	3.5	4.3
Italy	3.6	2.6	2.6	4.6	4.4	3.2
Japan	3.3	3.0	3.8	3.8	3.8	3.5
Average Ranks Across All Variables for Forecast Horizon h=3						
USA	2.6	4.2	2.8	3.4	5.0	3.0
UK	3.2	3.0	4.0	3.8	3.2	3.8
Germany	2.6	3.0	3.6	4.4	4.0	3.4
France	3.0	1.8	2.8	4.8	4.3	4.5
Italy	3.2	2.2	2.8	4.2	5.2	3.4
Japan	2.5	3.5	3.5	3.3	4.0	4.3
Average Ranks Across All Variables for Forecast Horizon h=4						
USA	2.2	4.0	3.2	3.6	4.2	3.8
UK	3.4	2.8	4.0	3.8	3.2	3.8
Germany	2.6	3.6	3.4	3.8	3.8	3.8
France	2.0	3.3	2.8	4.8	4.5	3.8
Italy	3.6	2.2	4.0	3.4	3.8	4.0
Japan	2.3	4.3	2.8	2.8	5.3	3.8

* Notes: See notes to Table 3. The experiment reported on in this table uses an updated version of the Global VAR dataset analyzed by Dees, di Mauro, Pesaran and Smith (2007). More specifically, quarterly forecasts are made for 5 variables including GDP (y), inflation (p), equity returns (q), short-term interest rates (ρ^s), and long-term interest rates (ρ^l). The ex ante forecasting period is 2021:Q4-2023:Q3, and predictions are made for 6 different countries and 4 different forecast horizons (i.e. h=1,2,3,4). Tabulated entries are average "ranks" of our different models, constructed by comparing MSFEs by country and forecast horizon, when averaged across all 5 variables. In our analysis, we compare 6 different models (i.e., PCA, CS, HT, PCA+macro, CS+macro, and HT+macro). All models include an AR component, and up to 2 factors estimated using PCA, with factors selected using the PC_{p2} statistic. PCA+macro, CS+macro, and HT+macro include lags of the 5 macroeconomic variables as explanatory variables, rather than just lags of the target variable, with lags selected using the SIC. Summarizing, model ranks are calculated for each variable (excluding deq for France and dp for Japan, due to data reliability issues for those two variables) and each forecast horizon, and are averaged across all five variables to yield the average ranks reported in the table. For complete details refer to Section 5.2